

日 本 国 特 許 庁

PATENT OFFICE
JAPANESE GOVERNMENT

JC912 U.S. PTO
09/714627
11/17/00

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日

Date of Application: 1999年11月19日

出 願 番 号

Application Number: 平成11年特許願第330236号

出 願 人

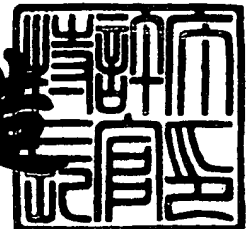
Applicant(s): 株式会社東芝

CERTIFIED COPY OF
PRIORITY DOCUMENT

2000年10月20日

特許庁長官
Commissioner,
Patent Office

及川耕造



【書類名】 特許願

【整理番号】 A009906696

【提出日】 平成11年11月19日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 15/00

【発明の名称】 構造化文書検索方法、構造化文書検索装置及び構造化文
書検索システム

【請求項の数】 14

【発明者】

 【住所又は居所】 神奈川県川崎市幸区柳町 7 0 番地 株式会社東芝柳町工
 場内

 【氏名】 服部 雅一

【発明者】

 【住所又は居所】 神奈川県川崎市幸区柳町 7 0 番地 株式会社東芝柳町工
 場内

 【氏名】 野々村 克彦

【発明者】

 【住所又は居所】 神奈川県川崎市幸区柳町 7 0 番地 株式会社東芝柳町工
 場内

 【氏名】 金輪 拓也

【特許出願人】

 【識別番号】 000003078

 【氏名又は名称】 株式会社 東芝

【代理人】

 【識別番号】 100058479

 【弁理士】

 【氏名又は名称】 鈴江 武彦

 【電話番号】 03-3502-3181

【選任した代理人】

【識別番号】 100084618

【弁理士】

【氏名又は名称】 村松 貞男

【選任した代理人】

【識別番号】 100068814

【弁理士】

【氏名又は名称】 坪井 淳

【選任した代理人】

【識別番号】 100092196

【弁理士】

【氏名又は名称】 橋本 良郎

【選任した代理人】

【識別番号】 100091351

【弁理士】

【氏名又は名称】 河野 哲

【選任した代理人】

【識別番号】 100088683

【弁理士】

【氏名又は名称】 中村 誠

【選任した代理人】

【識別番号】 100070437

【弁理士】

【氏名又は名称】 河井 将次

【手数料の表示】

【予納台帳番号】 011567

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1
【物件名】 要約書 1
【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 構造化文書検索方法、構造化文書検索装置及び構造化文書検索システム

【特許請求の範囲】

【請求項 1】

論理構造を持つ構造化文書データベースに対して、文書の論理構造を含む検索要求に基づいて検索を行う構造化文書検索方法であって、

前記検索要求に基づいて、文書の構造情報を含む検索グラフを生成し、

前記構造化文書データベースにおける実データに関するインデックス情報を利用して、前記検索グラフから、前記構造化文書データベースに対する検索処理手順を示す検索プランを生成し、

前記構造化文書データベースを検索対象として前記検索プランを実行することによって、前記検索要求を満足する検索結果を求めることを特徴とする構造化文書検索方法。

【請求項 2】

前記検索プランの生成においては、前記インデックス情報を利用しながら前記検索グラフを巡回することによって最適な検索プランを生成することを特徴とする請求項 1 に記載の構造化文書検索方法。

【請求項 3】

前記検索グラフ中において評価可能な部分グラフを優先的に評価する戦略に基づいて前記検索グラフを巡回することを特徴とする請求項に 2 記載の構造化文書検索方法。

【請求項 4】

前記検索プランの生成が全て完了した後に、該検索プランの実行を行うことを特徴とする請求項 1 に記載の構造化文書検索方法。

【請求項 5】

前記検索プランの生成および実行を交互に繰り返す行うことを特徴とする請求項 1 に記載の構造化文書検索方法。

【請求項 6】

前記構造化文書データベースは、要素名称および要素値に関する階層構造を含み、

前記検索要求は、前記要素名称および前記要素値に関する検索条件を含み、

前記インデックス情報は、前記構造化文書データベースにおける前記要素値の生起位置を特定する情報を含むデータ生起インデックスと前記構造化文書データベースにおける前記要素名称の生起位置を特定する情報を含む要素名称生起インデックスとの少なくとも一方を含むことを特徴とする請求項 1 に記載の構造化文書検索方法。

【請求項 7】

前記要素名称生起インデックスは、前記要素名称の生起位置を、前記要素名称の発生する部分構造の一階層上位の親要素によって指し示した情報を含むこと特徴とする請求項 6 に記載の構造化文書検索方法。

【請求項 8】

前記検索プランの生成においては、

ルール適用条件を示す情報と前記検索プランを構成すべき検索処理の内容を指示する情報とを含むプラン生成ルールが複数登録されたプラン生成ルールベースに基づき、プラン生成ルールを選択し、該プラン生成ルールを前記検索グラフの該当する要素に対して適用するとともに、該プラン生成ルールに含まれる検索処理を、前記検索プランを構成する 1 つの検索処理として決定し、

前記プラン生成ルールが適用された結果として影響が及ぶ前記検索グラフの要素に関して、プラン生成ルールの選択および適用ならびに前記検索プランにおいて後続させる検索処理の決定を行うことを、繰り返し行うことによって、

前記検索プランを生成していくことを特徴とする請求項 2 に記載の構造化文書検索方法。

【請求項 9】

前記プラン生成ルールには、前記インデックス情報を加味して決定されるコスト情報が付与されており、

前記コスト情報を考慮して、動的に、適用すべきプラン生成ルールを選択することを特徴とする請求項 8 に記載の構造化文書検索方法。

【請求項 1 0】

前記プラン生成ルールベースにおける前記プラン生成ルールを任意に登録および削除可能としたことを特徴とする請求項 8 に記載の構造化文書検索方法。

【請求項 1 1】

前記検索グラフの生成においては、前記検索要求の記述を構文解析した結果に基づいて前記検索グラフの生成を行うことを特徴とする請求項 1 に記載の構造化文書検索方法。

【請求項 1 2】

論理構造を持つ構造化文書データベースに対して、文書の論理構造を含む検索要求に基づいて検索を行う構造化文書検索装置であって、

前記検索要求に基づいて、文書の構造情報を含む検索グラフを生成する手段と

前記構造化文書データベースにおける実データに関するインデックス情報を利用して、前記検索グラフから、前記構造化文書データベースに対する検索処理手順を示す検索プランを生成する手段と、

前記構造化文書データベースを検索対象として前記検索プランを実行することによって、前記検索要求を満足する検索結果を求める手段とを備えたことを特徴とする構造化文書検索装置。

【請求項 1 3】

論理構造を持つ構造化文書データベースに対して、文書の論理構造を含む検索要求に基づいて検索を行うためのプログラムであって、

前記検索要求に基づいて、文書の構造情報を含む検索グラフを生成させ、

前記構造化文書データベースにおける実データに関するインデックス情報を利用して、前記検索グラフから、前記構造化文書データベースに対する検索処理手順を示す検索プランを生成させ、

前記構造化文書データベースを検索対象として前記検索プランを実行することによって、前記検索要求を満足する検索結果を求めるためのプログラムを記録したコンピュータ読取り可能な記録媒体。

【請求項 1 4】

論理構造を持つ構造化文書データベースに対して、文書の論理構造を含む検索要求に基づいて検索を行う構造化文書検索システムであって、

前記構造化文書データベースの実データを記憶する手段と、

前記構造化文書データベースにおける実データに関するインデックス情報を記憶する手段と、

外部から前記検索要求を受け付ける手段と、

受け付けた前記検索要求に基づいて、文書の構造情報を含む検索グラフを生成する手段と、

前記構造化文書データベースにおける実データに関するインデックス情報を利用して、前記検索グラフから、前記構造化文書データベースに対する検索処理手順を示す検索プランを生成する手段と、

前記構造化文書データベースを検索対象として前記検索プランを実行することによって、前記検索要求を満足する検索結果を求める手段と、

前記検索結果を外部へ出力する手段とを備えたことを特徴とする構造化文書検索システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、論理構造を持つ構造化文書データベースに対して、文書の論理構造を含む検索要求に基づいて検索を行う構造化文書検索方法、構造化文書検索装置及び構造化文書検索システムに関する。

【0002】

【従来の技術】

従来から文書データベースに対する検索要求を指定する方法としてキーワード指定がある。ユーザが検索要求をキーワード列という形式で文書データベースに要求すると、キーワード列を含んでいる文書群を返すというものである。

【0003】

このような素朴で原始的な検索要求方式は、全文検索エンジンなどに広く適用されているが、それゆえ、(1) 必要以上の文書群が検索されてしまうという低

い精度の問題や、(2) 利用部分以外のデータまで含んだ文書がデータ単位であるという粒度の問題がある。

【0004】

近年、SGML (Standard Generalized Markup Language) やXML (eXtensible Markup Language) などの構造化文書のための構造化文書規約が提案され、文書構造に基づいた検索要求の指定により、(1) 従来のキーワード検索よりも精度の高い検索と、(2) 利用部分だけのデータが得られるという木目細かい検索が可能になっている。しかしながら、この場合、予め文書構造を固定的なものに統一する必要がある、後から文書構造の変更ができない、あるいはデータ毎に文書構造を変えることができないという欠点がある。

【0005】

一方、RDB (Relational DataBase) では、表の構造に基づいた検索要求をSQL言語によって指定することができる。SQLは、ANSI X3, 1、およびISO/TC97/SC21/WG3 N117 (1987) において標準化されたRDBの問合せ言語である。しかしながら、文書構造はそのまま表形式に変換することは困難であり、RDBを文書データベースとしてそのまま用いることはできない。

【0006】

さらに、SGMLやXMLなどの構造化文書データベースに対するOODB (Object Oriented DataBase) で用いられた検索言語を適用する方法が考えられる。構造化文書は階層的な構造を持つため、各構成要素をオブジェクトとみなしたOODBと親和性が高いと考えられる。しかしながら、OODBでは、文書構造はあらかじめスキーマにより決定されていなければならない、子要素の任意繰り返しなど、オブジェクトモデルでモデル化するのは困難であり、オブジェクト指向データベースを文書データベースとしてそのまま用いることはできない。

【0007】

このような問題を解決するために、文書リポジトリに対して、SQLへ構造化

文書に適した拡張機能を追加した言語処理部を装備することが考えられている。構造化文書に適した拡張機能には、階層的な構造上の構成要素を特定するパス指定が第一に挙げられる。さらに、階層的な構造上の構成要素を特定するパスに正規表現などの曖昧性を含んだ曖昧パス指定や、階層的な構造のパターンを指定する構造パターン指定など、構造化文書が持つ構造的な揺らぎを吸収するような機能がSQLをベースに拡張されている。

【0008】

これらの特徴を持った検索要求を指定でき、かつ検索処理できる方式を提案しているものに、特開平6-203078号公報、特開平6-301721号公報、特開平11-15843号公報がある。

【0009】

特開平6-203078号公報（情報検索方法およびその装置）では、階層構造を全展開したパス集合を文字列表としてRDBに格納する方式を提案している。構造化文書を検索するとき、文字列表のパスを検索文の曖昧パスと文字列比較するSQLを発行することで、階層的な構造上の構成要素を特定している。この方式の問題点は、登録された文書数が増大すると、階層構造を全展開した文字列表が膨大なサイズになってしまうことである。

【0010】

特開平6-301721号公報（全文データベース検索方式）では、構成要素タイプをあらかじめ決めておき、その階層構造の親子関係や実データへのリンクなどを構成要素タイプ毎に構造情報としてRDB化する方法を提案している。構造化文書を検索するとき、検索要求をSQL文に変換している。この方式の問題点は、ルート要素から始まって、親要素から子供要素群へと展開して、階層的な構造上の構成要素を特定する検索処理方式であるため、登録された文書数が増大し階層木の深さと幅が増大すると、検索処理に要する計算量が膨大なものになってしまうことである。RDBの結合で展開処理を行っているため、実装システムは想像を超えた応答時間が予想される。特に曖昧パスが指定されたときは、その傾向が激しくなる。

【0011】

特開平 1 1 - 1 5 8 4 3 号公報（S G M L 文書検索装置および S G M L 文書検索方法）でも、構成要素タイプをあらかじめ決めておき、構成要素タイプ毎にデータを文字列結合した文書テーブルを作成しておく。構造化文書を検索するとき、検索要求を S Q L 文に変換している。この方式の問題点は、構成要素タイプ毎にデータをただ単に文字列結合するため、1 段レベルのパスしか指定できないことである。さらに、あらかじめ文書構造が決まっていなければならない、文章が持つ階層構造に対する柔軟な検索要求は発行できない、などの問題点も抱えている。

【0 0 1 2】

これらの方式では、データに対するインデックスと構造に関するインデックスを適切に組み合わせて、検索処理に要する計算量を抑えるようになっておらず、R D B のような最適化を入れにくい仕組みとなっている。

【0 0 1 3】

【発明が解決しようとする課題】

以上説明したように従来の技術では、（１）（曖昧パスを含む）文書が持つ階層構造に対する多様な検索指定を行うことと、（２）検索処理に要する計算量を膨大なものとしめないことというトレードオフの関係にある２つの要求を同時に満足させることは困難であった。

【0 0 1 4】

本発明は、上記事情を考慮してなされたもので、検索処理に要する計算量の増大を伴わずに、（曖昧パスを含む）文書が持つ階層構造に対する多様な検索指定を行うことを可能とした、構造化文書検索方法、構造化文書検索装置及び構造化文書検索システムを提供することを目的とする。

【0 0 1 5】

【課題を解決するための手段】

本発明（請求項 1）は、論理構造を持つ構造化文書データベースに対して、文書の論理構造を含む検索要求に基づいて検索を行う構造化文書検索方法であって、前記検索要求に基づいて、文書の構造情報を含む検索グラフを生成し、前記構造化文書データベースにおける実データに関するインデックス情報を利用して、

前記検索グラフから、前記構造化文書データベースに対する検索処理手順を示す検索プランを生成し、前記構造化文書データベースを検索対象として前記検索プランを実行することによって、前記検索要求を満足する検索結果を求めることを特徴とする。

【0016】

好ましくは、前記検索プランの生成においては、前記インデックス情報を利用しながら前記検索グラフを巡回することによって最適な検索プランを生成するようにしてもよい。

【0017】

好ましくは、前記検索グラフ中において評価可能な部分グラフを優先的に評価する戦略に基づいて前記検索グラフを巡回するようにしてもよい。

【0018】

好ましくは、前記検索プランの生成が全て完了した後に、該検索プランの実行を行うようにしてもよい。

【0019】

好ましくは、前記検索プランの生成および実行を交互に繰り返し行うようにしてもよい。

【0020】

好ましくは、前記構造化文書データベースは、要素名称および要素値に関する階層構造を含み、前記検索要求は、前記要素名称および前記要素値に関する検索条件を含み、前記インデックス情報は、前記構造化文書データベースにおける前記要素値の生起位置を特定する情報を含むデータ生起インデックスと前記構造化文書データベースにおける前記要素名称の生起位置を特定する情報を含む要素名称生起インデックスとの少なくとも一方を含むようにしてもよい。

【0021】

好ましくは、前記要素名称生起インデックスは、前記要素名称の生起位置を、前記要素名称の発生する部分構造の一階層上位の親要素によって指し示した情報を含むようにしてもよい。

【0022】

好ましくは、前記検索プランの生成においては、ルール適用条件を示す情報と前記検索プランを構成すべき検索処理の内容を指示する情報とを含むプラン生成ルールが複数登録されたプラン生成ルールベースに基づき、プラン生成ルールを選択し、該プラン生成ルールを前記検索グラフの該当する要素に対して適用するとともに、該プラン生成ルールに含まれる検索処理を、前記検索プランを構成する1つの検索処理として決定し、前記プラン生成ルールが適用された結果として影響が及ぶ前記検索グラフの要素に関して、プラン生成ルールの選択および適用ならびに前記検索プランにおいて後続させる検索処理の決定を行うことを、繰り返し（伝播的に）行うことによって、前記検索プランを生成していくようにしてもよい。

【0023】

好ましくは、前記プラン生成ルールには、前記インデックス情報を加味して決定されるコスト情報が付与されており、前記コスト情報を考慮して、動的に、適用すべきプラン生成ルールを選択するようにしてもよい。

【0024】

好ましくは、前記プラン生成ルールベースにおける前記プラン生成ルールを任意に登録および削除可能とするようにしてもよい。これによって、検索プランの生成をカスタマイズすることができる。

【0025】

好ましくは、前記検索グラフの生成においては、前記検索要求の記述を構文解析した結果に基づいて前記検索グラフの生成を行うようにしてもよい。

【0026】

また、本発明（請求項12）は、論理構造を持つ構造化文書データベースに対して、文書の論理構造を含む検索要求に基づいて検索を行う構造化文書検索装置であって、前記検索要求に基づいて、文書の構造情報を含む検索グラフを生成する手段と、前記構造化文書データベースにおける実データに関するインデックス情報を利用して、前記検索グラフから、前記構造化文書データベースに対する検索処理手順を示す検索プランを生成する手段と、前記構造化文書データベースを検索対象として前記検索プランを実行することによって、前記検索要求を満足す

る検索結果を求める手段とを備えたことを特徴とする。

【0027】

また、本発明（請求項13）は、論理構造を持つ構造化文書データベースに対して、文書の論理構造を含む検索要求に基づいて検索を行うためのプログラムであって、前記検索要求に基づいて、文書の構造情報を含む検索グラフを生成させ、前記構造化文書データベースにおける実データに関するインデックス情報を利用して、前記検索グラフから、前記構造化文書データベースに対する検索処理手順を示す検索プランを生成させ、前記構造化文書データベースを検索対象として前記検索プランを実行することによって、前記検索要求を満足する検索結果を求めるさせるためのプログラムを記録したコンピュータ読取り可能な記録媒体である。

【0028】

また、本発明（請求項14）は、論理構造を持つ構造化文書データベースに対して、文書の論理構造を含む検索要求に基づいて検索を行う構造化文書検索システムであって、前記構造化文書データベースの実データを記憶する手段と、前記構造化文書データベースにおける実データに関するインデックス情報を記憶する手段と、外部から前記検索要求を受け付ける手段と、受け付けた前記検索要求に基づいて、文書の構造情報を含む検索グラフを生成する手段と、前記構造化文書データベースにおける実データに関するインデックス情報を利用して、前記検索グラフから、前記構造化文書データベースに対する検索処理手順を示す検索プランを生成する手段と、前記構造化文書データベースを検索対象として前記検索プランを実行することによって、前記検索要求を満足する検索結果を求める手段と、前記検索結果を外部へ出力する手段とを備えたことを特徴とする。

【0029】

なお、方法に係る本発明は装置／システムに係る発明としても成立し、装置／システムに係る本発明は方法に係る発明としても成立する。

【0030】

また、方法または装置／システムに係る本発明は、コンピュータに当該発明に相当する手順を実行させるための（あるいはコンピュータを当該発明に相当する

手段として機能させるための、あるいはコンピュータに当該発明に相当する機能を実現させるための) プログラムを記録したコンピュータ読取り可能な記録媒体としても成立する。

【0031】

本発明では、必要に応じて(要素名称生起やデータ生起など)様々なインデックス情報や文書の階層構造に関する情報を有効に利用して、(文書の構造情報を含んだ)検索グラフを最適に巡回することで最適な検索プランを生成し、これを実行する。すなわち、本発明によれば、存在するインデックス情報を動的に用いながら、最適な検索プラン生成・実行することが可能になる。また、本発明によれば、(構造照合やデータ比較を組み合わせた)多様な検索指定が行われていても、最適な検索プランを生成することで、検索処理に要する計算量を抑えることができる。

【0032】

このように本発明によれば、(1)(曖昧パスを含む)文書が持つ階層構造に対する多様な検索指定を行いながら、(2)検索処理に要する計算量を膨大なものとし、という両要求を同時に満足し、論理構造を持った構造化文書データベースに対して文書の論理構造を含めた検索要求文で検索するサービスを実現することができる。

【0033】

【発明の実施の形態】

以下、図面を参照しながら発明の実施の形態を説明する。

【0034】

本発明を適用可能な構造化文章として、例えば、SGML(Standard Generalized Markup Language)やXML(eXtensible Markup Language)で記述された文書が挙げられる。SGMLとは、ISO(国際標準機構)で定められた規格である。XMLとは、W3C(ワールドワイドウェブコンソーシアム)にて定められた規格である。それぞれ文書を構造化することを可能とする構造化文書規約である。

【0035】

SGMLやXMLを用いた文書の構造の表現にはタグが用いられる。タグには、開始タグと終了タグがあり、文書構造情報の構成要素を開始タグと終了タグで囲むことにより、文書中の文章の区切りと、その文書が構造上どの構成要素に属するのかとを明確にする。ここで、開始タグは「要素名称」を記号「<」と「>」で閉じたものであり、終了タグは「要素名称」を記号記号「<」と「/>」で閉じたものである。タグに続く構成要素の内容が、テキストまたは子供の構成要素の繰り返しである。また、開始タグには「<要素名称 属性=“属性値”>」のように属性情報を設定することができる。

【0036】

以下では具体例としてXMLを用いて説明するものとする。

【0037】

また、データベースの内容の具体例として特許出願に関する情報を用い、検索の具体例として特許出願に関する情報の検索を用いるものとする。なお、具体例を用いた説明で「特許」という場合は、「特許出願に関する(もの)」というような意味で用いているものとする。

【0038】

図1に、本発明の一実施形態に係る構造化文書データベース・システムのシステム構成を示す。

【0039】

本システムは、要求制御部1、格納処理部2、検索処理部3、データファイル4、インデックスファイル5を含む。

【0040】

本システム構成は、ソフトウェアを用いて実現可能である。なお、データファイル4、インデックスファイル5は、例えば外部記憶装置を用いて構成される。

【0041】

要求制御部1は、ユーザからの検索要求や格納要求など構造化文書データベースへの要求を処理し、検索処理部3や格納処理部2へ処理を渡す処理部である。検索要求と格納要求は要求受付部11でメッセージとして受け取る。受け取ったメッセージについて要求処理部12で検索要求か格納要求かの分別を行い、検索

処理部 3 による検索処理あるいは格納処理部 2 による格納処理を呼び出す。また、検索処理部 3 から渡された検索結果は、結果処理部 1 3 にて整形されて、要求元のユーザに返される。

【 0 0 4 2 】

検索処理部 3 は、検索要求を解析し、検索要求を満足する検索結果を生成する処理部である。検索要求構文解析部 3 1 にて、検索要求から字句切り出しや要求文の構造抽出を行い、検索グラフ生成部 3 2 にて、検索グラフを生成する。検索プラン生成部 3 3 にて、生成された検索グラフから検索プランを生成し、検索プラン実行部 3 4 にて、生成された検索プランを実行し、検索要求を満足する検索結果を生成する。検索結果は、要求制御部 1 に渡される。

【 0 0 4 3 】

格納処理部 2 は、格納要求を解析し、構造化文書を格納する処理部である。格納要求構文解析部 2 1 にて、構造化文書から字句切り出しや構造化文書の構造抽出を行う。データ格納部 2 2 にて、構造化文書のデータや構造データをデータファイル 4 に格納し、インデックス格納部 2 3 にて、構造化文書のデータや構造データに対するインデックスをインデックスファイル 5 に格納する。なお、インデックスファイルの作成・更新は、格納すべき構造化文書が入力されるごとに行ってもよいし、適宜まとめて行ってもよい（検索の効率化のためには、前者の方が好ましい）。

【 0 0 4 4 】

図 2 に、構造化文書の一例を示す。

【 0 0 4 5 】

図 2 は、構造化文書の一例として「特許」情報の例を示したものである（XML で記述した例である）。

【 0 0 4 6 】

「特許」タグ（すなわち、＜特許＞と＜／特許＞の対；他も同様の意味である）で囲まれた内部には、「名称」タグで囲まれた「名称」情報、「出願人」タグで囲まれた「出願人」情報、「出願番号」タグで囲まれた「出願番号」情報、「出願日」タグで囲まれた「出願日」情報、「要約」タグで囲まれた「要約」情報

、「キーワード」タグで囲まれた「キーワード」情報が存在する。

【0047】

「出願日」情報は、さらに、「年」タグで囲まれた「年」情報、「月」タグで囲まれた「月」情報、「日」タグで囲まれた「日」情報により構成される。なお、「出願日」情報は、「年号」情報をさらに含んでもよい。あるいは、「年」情報を西暦で表してもよい。

【0048】

また、「キーワード」情報としては、1または複数個のものを指定することができる（図2の例では2個のキーワード「XML」、「検索」が指定されている）。

【0049】

この「キーワード」情報のように、XMLなどの構造化文書では、任意の構成要素の繰り返しを含んでいたりと、さらには文書構造があらかじめ決まっていない（RDBやOODBのスキーマ定義では定義できない）のが通常である。

【0050】

なお、「特許」情報には、「公開番号」情報や、「特許番号」情報、あるいはその他の種々の情報を含めることができる。

【0051】

図3および図4に、本実施形態で必要に応じて検索で使用する概念階層を構造化文書で表現した例を示す。図3および図4の例は、「概念」情報をXMLで記述したものである。

【0052】

図3の「概念」情報の例は、いわゆる特許調査における特許文書の内容を分類するための一つの分類軸として用いる「情報モデル」を概念階層で表現している。「概念」タグで囲まれた「概念」情報は、入れ子構造を持った文書構造を持っている。つまり、図3の例では、概念「情報モデル」の子概念として、概念「ドキュメント」、概念「リレーション」、概念「オブジェクト」が存在している。また、概念「ドキュメント」の子概念として、概念「構造化ドキュメント」、概念「非構造化ドキュメント」が存在し、さらに、概念「構造化ドキュメント」

」の子供概念として、概念「XML」、概念「SGML」が存在している。

【0053】

図4の概念階層の記述例は、図3とは異なる分類軸「情報操作」を概念階層で表現している。図4の例では、概念「情報操作」の子供概念として、概念「検索」、概念「格納」、概念「加工」、概念「流通」が存在している。

【0054】

図5に、本実施形態における構造化文書データベースの概念的な構造の例を示す。

【0055】

構造化文書を集めた構造化文書データベースは、例えばUNIXのディレクトリ構造のように階層的に格納されていることを指定している。

【0056】

構造化文書データベースの階層木の各ノード（図5では番号が付され円形で示されたもの）を、ドキュメントノードと呼ぶ。なお、以下では、ドキュメントノードをDノードと呼ぶ。

【0057】

任意のDノード以下の部分階層木は、構造化文書データベースから切り出された構造化文書を示している。

【0058】

各Dノードには、オブジェクトID（図5では円内部に記述されたもの）が割り当てられる。オブジェクトIDは、構造化文書データベース内ではユニークな数値を持つものとする。

【0059】

図5の例では、階層木のルートとなるドキュメントノード（根Dノード）に、それが根Dノードであることを特定可能なオブジェクトID「#0」が割り当てられるものとしている。

【0060】

図5の例において、根Dノードすなわち「#0」のDノードからは、「root」タグを先頭に持つ「#17」のDノードへリンクが張られている。「#17

」Dノードからは、「IR特許」タグを先頭に持つ「#21」Dノード、「DB特許」タグを先頭に持つ「#45」Dノード、「概念」タグを先頭に持つ「#78」Dノードへのリンクがそれぞれ張られている。なお、IR特許とは、例えば、IR技術に係る発明をその明細書中に含む特許出願というような意味である（DB特許、OODB特許、RDB特許についても同様である）。

【0061】

図2に例示された「特許」情報は、「#902」のDノード以下の部分階層木に対応しており、「名称」タグあるいは「キーワードタグ」などを先頭に持つ各々の末端のDノード（#903～#905、#907～#912）からは、「情報検索装置」、「T社」、「特願平10-××××××」、「10」、「3」、「12」、「情報の提示形式の変更が～（以下、省略）」、「XML」、「検索」などの文字列（要素値）へのリンクがそれぞれ張られている。

【0062】

ところで、「#639」のDノード以下の部分階層木も一つの「特許」情報に対応する部分であるが、根Dノードからみて「#902」Dノードと「#639」Dノードとは階層の深さが異なっている。このように、根Dノードから「特許」情報に対応するDノードまでの階層関係は任意に設定することが可能である。

【0063】

すなわち、図5に示されているように「特許」情報は、「#902」Dノードや「#639」Dノードなどのように階層木上の任意の部分に発生し得る。これが構造化文書データベースの特徴である。そのため、階層木上の任意の部分に発生した「特許」情報を検索したいという検索要求がある。

【0064】

なお、本実施形態では、図5に示すように、図3や図4のような「概念」情報も構造化文書データベース内に併せて保持することができる（例えば、「#78」Dノード下位の以下の部分階層木に含まれる）。

【0065】

図6に、本実施形態における構造化文書データベースへの構造化文書の格納コマンドの一例を示す。

【0066】

コマンド名「Insert」の後に、格納先「“root/IR特許”」、格納データ「“<特許>～(中略)～</特許>”」の2つのパラメータが存在する。この記述は、格納先として、「root」タグを先頭に持つ部分階層木から辿って、「IR特許」タグを先頭に持つ部分階層木の先頭要素に、格納データ「“<特許>～(中略)～</特許>”」を挿入することを意味する。「“root/IR特許”」を文書パスと呼ぶ。

【0067】

図6に例示した格納コマンドを実行した結果として、図5に例示した概念的な構造の「#902」のDノード以下の部分階層木が作られることになる。

【0068】

「Insert」コマンド名を持つ格納要求は、図1の要求制御部1にて受理され、格納処理部2による構文解析(21)を経て、データ格納(22)とインデックス格納(23)が行われる。

【0069】

図7に、構造化文書データベースへの検索コマンドの一例を示す。

【0070】

図7の例は、検索コマンドをSQLに似たSelect文で表現したもので、『構造化文書データベース中に出現する「特許」情報のうちその「キーワード」情報として「検索」を持つものについて、「出願番号」情報を抽出し、それを「文献」情報として出力せよ』という検索要求を意味している。

【0071】

「Where」句が条件部分を示しており、「From」句が文書パス指定部分を示しており、「Select」句が情報抽出部分を示している。「\$1」、「\$2」はデータが束縛される変数である。

【0072】

「<*/特許>」のように要素名称の前が「*」で始まっているならば、指定された文書パスの任意子孫の「特許」にマッチすることができる。「root/*/特許」のように曖昧な文書パスが、『階層木上の任意の部分に発生した「特許」

情報を検索したいという検索要求』に対応する。

【0073】

例えば、図5において、「#902」Dノード以下の部分階層木に対応する「特許」情報について、「#912」Dノードからリンクされた「検索」が条件を満たし、「#905」Dノードからリンクされた「特願平10-XXXXXX」が検索結果となる。

【0074】

図8に、構造化文書データベースへの検索コマンドの他の例を示す。

【0075】

この例は、『構造化文書データベース中に出現する「特許」情報のうち、その「キーワード」情報として、概念「ドキュメント」に属する内容（図3では、概念の名前の属性値（文字列）に一致する要素値を持つ「特許」情報について、「出願番号」情報を抽出し、それを「文献」情報として出力せよ』という検索要求を意味している。

【0076】

この例では、「特許」情報と「概念」情報の2つを参照し、それぞれ「キーワード」と「名前」に対して、同一変数「\$x2」が割り当てられている。これは2つの情報の結合処理を意味している。

【0077】

例えば、図5において、「概念」情報が図3のようであるとすると、「#911」のDノードからリンクされた「XML」が図3のように概念「ドキュメント」に属するので、「#905」のDノードからリンクされた「特願平10-XXXXXX」が検索結果となる。

【0078】

図9に、構造化文書データベースへの検索コマンドのさらに他の例を示す。

【0079】

この例は、『構造化文書データベース中に出現する「特許」情報に対して、概念「情報モデル」での分類と概念「情報操作」での分類の2分類軸を設定して、「出願番号」と「情報モデル」軸と「情報操作」軸とを抽出して「文献」情報と

して検索せよ』を意味している。

【0080】

この例では、「特許」情報と「概念」情報の2つを参照し、それぞれ「キーワード」と「名前」に対して、同一変数「\$ x 2」が割り当てられている。これも2つの情報の結合処理を意味している。

【0081】

「情報モデル」軸を取り出す部分では、「特許」情報の「キーワード」情報「\$ x 2」が文書パス「root」以下の概念「情報モデル」の任意子孫の「概念」情報にマッチするものを探索し、概念「情報モデル」の1つ子供の概念に置き換えて「\$ x 3」とする処理が組み込まれている。「情報操作」軸を取り出す部分も同様に、「特許」情報の「キーワード」情報「\$ x 2」が文書パス「root」以下の概念「情報操作」の任意子孫の「概念」情報にマッチするものを探索し、概念「情報操作」の1つ子供の概念に置き換えて「\$ x 4」とする処理が組み込まれている。

【0082】

例えば、図5において、「概念」情報が図3および図4のようであるとすると、「# 9 1 1」のDノードからリンクされた「XML」が図3のように概念「情報モデル」に属し、かつ、「# 9 1 2」のDノードからリンクされた「検索」が図4のように概念「情報操作」に属するので、「# 9 0 5」のDノードからリンクされた「特願平 1 0 - x x x x x x」と、図3の概念「情報モデル」の1つ子供の概念「ドキュメント」と、図4のように概念「情報操作」の1つ子供の概念「検索」とが検索結果となる。

【0083】

図10に、図9の検索要求を処理した検索結果の一例を示す。図10に例示されるように、検索結果もXMLで表現することができる。

【0084】

図9で示された検索要求は、図1の要求制御部1にて受理され、検索処理部3にて構文解析(31)、検索グラフ生成(32)、検索プラン生成(33)、検索プラン実行(34)などの一連の処理を経て、要求制御部1の結果処理部13

にて整形されて、図 1 0 に示すような検索結果が得られる。

【 0 0 8 5 】

先にも述べたように、「特許」情報に対して、概念「情報モデル」での分類と概念「情報操作」での分類の 2 分類軸を設定して、「出願番号」情報とともにまとめられて「文献」情報のリストとして表示されている。例えば、第一の「文献」情報では、『「特願平 1 0 - × × × × × ×」の特許が「ドキュメント」×「検索」で分類されている』ことを意味している。

【 0 0 8 6 】

以下では、検索処理部 3 における処理についてより詳しく説明する。

【 0 0 8 7 】

図 1 1 および図 1 2 に、図 9 の検索要求に対して検索グラフ生成部 3 2 が生成する検索グラフの一例を示す（なお、図 1 1、図 1 2 は便宜上、同一の検索グラフの一部を省略したものであって、すなわち、図 1 1 は同一の検索グラフの C O N 以下の部分を省略したものであり、図 1 2 は同一の検索グラフの A N D 以下の部分を省略したものである）。

【 0 0 8 8 】

図 1 1 および図 1 2 に示されるように、検索グラフは、双方向リンク（図中の両方向の矢印）とノード（図中の円形、四角形、六角形）を含むネットワークを形成する。

【 0 0 8 9 】

図 1 1 および図 1 2 において、四角形で示されるノードは、具体的なデータ（文字列）を表している。四角形で示されるノードを除く各ノードを、検索グラフノード（以下、G ノード）と呼ぶ。すなわち、G ノードは、六角形で示される G ノードと円形で示される G ノードの 2 種類から構成される。

【 0 0 9 0 】

円形で示される G ノードは、変数を表す G ノードであり、「\$ __」で始まる文字列を持っている。変数を表す G ノードは、内部的に生成された変数と、それ以外の「\$ x 1」など検索要求の S e l e c t 文に含まれている変数とに分類できる。

【0091】

六角形で示されるGノードは、「QUERY」のGノード、「AND」のGノード、「TAG」のGノード、「ATT」のGノード、「VAL」のGノードなどから構成されている。

【0092】

「Query」Gノードは図9のような検索要求のSelect文全体に対応しており、「AND」GノードはWhere句に対応しており、「CON」GノードはSelect句に対応している。Where句以下の複合的な条件部分は、「AND」Gノードと「AND」Gノードから出ているサブネット群に対応している。

【0093】

一例として、『指定された文書パス「root」以下の任意の「特許」情報』の条件は、3つの「TAG」Gノード（「root」、「*」、「特許」を持つGノード列）で表現されている。2つの「TAG」Gノードを繋ぐ変数Gノード（例えば、「\$__1」や「\$__2」）は、図5で示されるDノードで束縛可能な変数である。例えば、「\$__2」変数のGノードは、右側2つの「TAG」Gノードから解釈すると『指定された文書パス「root」以下の任意の文書』を表し、左の「TAG」Gノードと接続しているため、それも併せて解釈すると『指定された文書パス「root」以下で「特許」タグを先頭に持つ任意の文書』を表す。

【0094】

「ATT」Gノード、「VAL」Gノードは、それぞれ属性、要素データの関係を示している。

【0095】

また、六角形で示されるGノードには複数のリンクが接続する。

【0096】

「QUERY」Gノードには、「AND」に接続するop1リンク、「CON」に接続するop2リンクがある。

【0097】

「AND」Gノードには、「QUERY」に接続するop1リンク、「TAG」群に接続するop2リンク群がある。

【0098】

「TAG」Gノードには、上位Gノードに接続するop1リンク（左側）、データに接続するop2リンク（下側）、下位Gノードに接続するop3リンク（右側）がある。

【0099】

「ATT」Gノードには、上位Gノードに接続するop1リンク（上側）、データに接続するop2リンク（右側）、下位Gノードに接続するop3リンク（下側）がある。

【0100】

「CON」Gノードには、「QUERY」に接続するop1リンク、「TAG」に接続するop2リンクがある。

【0101】

「VAL」Gノードには、上位Gノードに接続するop1リンク（左側）、下位Gノードに接続するop2リンク（下側）がある。

【0102】

また、変数を表すGノード（円形で示されるGノード）には、他のGノード群に接続するopリング群がある。

【0103】

図9の例では、前述したように、「特許」情報と「概念」情報の2つを参照し、それぞれ「キーワード」と「名前」に対して、同一変数「\$x2」が割り当てられている。「\$x2」の変数Gノードは、3つの「VAL」Gノードへのopリンクとして接続し、逆に3つの「VAL」Gノードからop2リンクとして接続されている。

【0104】

また、Select句に対応している「CON」Gノードより下位のGノードがネットワークを形成している。「特許」情報は、「出願番号」情報、「軸」属性が「情報モデル」の「分類」情報、「軸」属性が「情報操作」の「分類」情報

から構成されている。「\$x1」、「\$x3」、「\$x4」などの変数Gノードは「AND」GノードであるWhere句が処理された後に変数値が確定し、束縛されて、図10に示す結果となる。

【0105】

なお、図11および図12に例示された検索グラフは、例えばYacc (Yet Another Compiler-Compiler) / Lex (a LEXical analyzer generator) などの既存の構文解析プログラムジェネレータに、図9のような検索要求の記述を入力することによって生成することができる。

【0106】

次に、本実施形態で用いるインデックスファイルについて説明する。

【0107】

図13に、インデックスファイルの一種である要素名称生起インデックスの概念的な構造の一例を示す。

【0108】

要素名称生起インデックスとは、構造化文書データベースに格納されている要素名称のリストと、各要素名称が先頭に発生する構造化文書の位置とを関連付けてインデックスファイル化したものである。

【0109】

例えば図5の構造化文書データベースのように、(「特許」情報に対応する)「特許」という要素名称が、Dノード群「#902」、「#639」、…により示される構造化文書において発生している場合、これをインデックス化すると、図13に示すように、Dノード群「#902」、「#639」、…の親Dノード「#21」、「#67」、…が要素名称生起インデックスファイルに「特許」キーからのチェーンで格納される。

【0110】

このように親Dノードでインデックス化すると、インデックスファイルサイズを圧縮することができる。すなわち、親Dノードでインデックスすれば、子Dノードが増大しようとも、親Dノードで代用しているのでチェーンサイズは増大し

ない。これに対して、実Dノードをインデックス化すれば「特許」文書の格納数の増大とともにチェーンサイズはそれに比例して増加してしまう

図 1 4 に、インデックスファイルの一種であるデータ生起インデックスの概念的な構造の一例を示す。

【0 1 1 1】

データ生起インデックスとは、構造化文書データベースに格納されている文字列データのリストと各文字列データが発生する構造化文書の位置とを関連付けてインデックスファイル化したものである。

【0 1 1 2】

例えば図 5 の構造化文書データベースのように、「検索」という文字列データが、Dノード群「# 9 1 2」、「# 6 4 7」、「# 6 5 0」、…により示される構造化文書にて発生している場合（なお、「# 6 4 7」のDノードからリンクされるデータ中に検索という文字列が含まれているものとする）、これをインデックス化すると、図 1 4 に示すように、Dノード群「# 9 1 2」、「# 6 4 7」、「# 6 5 0」…がデータ生起インデックスファイルに「検索」キーからのチェーンで格納される。

【0 1 1 3】

なお、逆階層インデックスなど、その他のインデックスファイルを用いてもよい。逆階層インデックスとは、あるノードとその親ノードとの対応を格納したものである（あるノードからその親ノードを求めることができる）。

【0 1 1 4】

次に、本実施形態の構造化文書データベースの検索プラン生成部 3 3 について説明する。

【0 1 1 5】

図 1 5 に、検索プラン生成部 3 3 の構成例を示す。図 1 5 は、検索グラフ生成部 3 2 にて生成された検索グラフを入力として実行プランリストを出力する検索プラン生成部 3 3 の構成を表している。

【0 1 1 6】

図中 6 は、後述するプラン生成ルールを格納したプラン生成ルール格納部であ

る（なお、プラン生成ルール格納部 6 は例えば外部記憶装置を用いて構成される）。

【0 1 1 7】

候補 G ノード登録部 3 3 1 は、図 1 1 および図 1 2 に例示したような検索グラフを構成する各 G ノードを、候補 G ノードリストへ登録する。

【0 1 1 8】

G ノードルール発火チェック部 3 3 2 は、候補 G ノードリストを構成する G ノードに対してプラン生成ルールの適用をチェックする。

【0 1 1 9】

プラン生成ルール適用部 3 3 3 は、プラン生成ルールの適用可能な各 G ノードについて、コスト最小の G ノードとプラン生成ルールとのペアを取り出し、プラン生成ルールを実行する。プラン生成ルールを実行した結果である実行プランが、実行プランリスト 3 3 5 へ追加される。

【0 1 2 0】

また、プラン生成ルールによって値が具体化される可能性のある変数 G ノード群について、候補 G ノードリスト 3 3 4 に登録する。候補 G ノードリスト 3 3 4 が空になるまで、これを繰り返す。

【0 1 2 1】

このように、プラン生成ルールベースを使って、検索グラフの各要素に対してプラン生成ルールを適用し、適用された結果として影響のある検索グラフの各要素に対して再度プラン生成ルールを適用すること、すなわち、検索グラフを伝播的に巡回することで、格段に効率的な検索プランを実現することができる。

【0 1 2 2】

なお、全ての検索プランの生成が完了した後に、生成された検索プランを実行するようにしてもよい。また、1つの検索プランの生成とその検索プランの実行とを一纏まりとして続けて行い、これを繰り返し実行する（すなわち、検索プランの生成とその実行を交互に繰り返し行う）ようにしてもよい。

【0 1 2 3】

図 1 6 に、プラン生成部 3 3 で利用するプラン生成ルールの一例を示す。

【0 1 2 4】

図 1 6 のプラン生成ルール例は、1 1 個のルールをテーブル形式で記述したものである。

【0 1 2 5】

各ルールには、ルール番号、適用可能な G ノードのクラス、適用コスト、適用条件 (I F) 部、アクション (T H E N) 部の属性が存在する。

【0 1 2 6】

コストは、0 以上 1 以下の f l o a t 値を持ち、大きい数値を持つほど計算コストが大きくなることを意味するものとする。

【0 1 2 7】

適用条件 (I F) 部においては、O P 1 ~ 3 は前述したリンクを表す。また、図 1 6 中の「具」はそのリンクの変数 G ノードが具体化されていることを表し、「未」は具体化されていないことを表し、「*」はそのリンクの先のデータが「*」であることを表し、「AND」はそのリンク先が「AND」G ノードであることを表す。なお、変数 G ノードが具体化されているとは、『変数 G ノードが取りうる値が枚挙可能な状態である』ことと定義する。

【0 1 2 8】

適用条件 (I F) 部の「その他」の部分は、その他の適用条件を示す。例えば、ルール番号「0 3」における「O P 2 に要素名称生起インデックスが存在」は、その O P 2 の具体化されている変数値と一致する要素名称が要素名称生起インデックスに存在することを適用条件とするものである。

【0 1 2 9】

アクション (T H E N) 部のオペレータは、詳しくは後述するように検索プラン実行部 3 4 で実行されるアクションを示す。

【0 1 3 0】

図 1 6 において、例えば、ルール番号「0 1」は、「T A G」G ノードに対して適用可能で、コストが 1. 0 であることを示している。さらに、適用条件が『o p 1 リンクの変数 G ノードが「AND」であり、o p 2 リンクの変数 G ノード (含むデータ) が具体化されていて、o p 3 リンクの変数 G ノードが具体化され

ていない』ことである。またアクションが、『実行プランPATHINSTを生成する』ことを示している。

【0131】

また、例えば、ルール番号「02」は、「TAG」Gノードに対して適用可能で、コストが0.5であることを示している。さらに、適用条件が『op1リンクの変数Gノードが具体化されていて、op2リンクの変数Gノード（含むデータ）が具体化されていて、op3リンクの変数Gノードが具体化されていない』ことである。またアクションが、『実行プランPATHEXPAND1を生成する』ことを示している。

【0132】

図17に、検索プラン実行部34で利用されるオペレータの一例を示す。

【0133】

検索プラン実行部34では、入力された実行プランリストを1つずつ取り出す（フェッチする）処理341と、実行する処理342とを繰り返し、その結果を検索結果として出力とする。

【0134】

各オペレータの処理内容は次の通りである。

【0135】

- ①PATHINST : 文書パス「root」を取り出す処理。
- ②PATHEXPAND1 : 指定された要素名称をキーにして、上位Dノード群からキーにマッチするDノード群を算出する処理。
- ③PATHEXPAND2 : インデックス化されている要素名称をキーにして、構造化文書データベース内で発生する親子Dノード群を算出する処理。
- ④PATHEXPAND3 : インデックス化されている要素名称をキーにして、子供Dノードから親Dノードを算出する処理。
- ⑤PATHCHECK : 2つのDノード集合が与えられたとき、それらが指定された要素名称で親子関係にある2つのDノードの組み合わせを算出する処理。
- ⑥JOIN : 変数Gノードxがopリンクで接続している複数のGノード

から具体化が進行して、xで重なり合ったときに行われる結合演算処理。

⑦VALUE : 変数Gノードxの要素データの候補を算出する処理。

⑧SELECT : 変数Gノードxに対する要素データを選択するときの比較演算処理。

⑨FIND : インデックス化されている要素データの候補を算出する処理。

【0 1 3 6】

図 1 8 に、検索プラン生成部 3 3 の処理手順の一例を示す。

【0 1 3 7】

まず、候補Gノードリスト 3 3 4 と実行プランリスト 3 3 5 を空リストとして初期化する（ステップ S 1）。

【0 1 3 8】

検索グラフを構成するGノード全部を候補Gノードリスト 3 3 4 に登録する（ステップ S 2）。

【0 1 3 9】

中間変数 r s e t を空リストとして初期化する（ステップ S 3）。

【0 1 4 0】

候補Gノードリスト 3 3 4 が空リストであれば（ステップ S 4）、検索プラン生成部 3 3 を終了する（ステップ S 4 1）。

【0 1 4 1】

候補Gノードリスト 3 3 4 が空リストでなければ（ステップ S 4）、空リストでない候補Gノードリスト 3 3 4 の各々の構成要素xに対して、ステップ 5 1 からステップ 5 4 まで繰り返す（ステップ S 5）。

【0 1 4 2】

構成要素xに適用可能なプラン生成ルール群を検索する（ステップ S 5 1）。

【0 1 4 3】

検索されたプラン生成ルール群から I F 部を満足するプラン生成ルール群を選択する（ステップ S 5 2）。

【0 1 4 4】

選択されたプラン生成ルール群がなければ（ステップ S 5 3）、候補 G ノードリスト 3 3 4 から構成要素 x を削除する（ステップ S 5 3 1）。

【0 1 4 5】

選択されたプラン生成ルール群があれば（ステップ S 5 3）、各々のプラン生成ルール r に対してステップ S 5 4 1 を適用する（ステップ S 5 4）。

【0 1 4 6】

プラン生成ルール r のコスト c を計算し、 $r\ set$ に $\langle x, r, c \rangle$ を追加する（ステップ S 5 4 1）。

【0 1 4 7】

続いて、 $r\ set$ の各要素 $\langle x, r, c \rangle$ に対して、最小のコスト c を持つ要素 $\langle x_1, r_1, c_1 \rangle$ を選択する（ステップ S 6）。ここで、実行プランリスト 3 3 5 に、所定の事項を登録する。

【0 1 4 8】

候補 G ノードリスト 3 3 4 から構成要素 x_1 を削除する（ステップ S 7）。

【0 1 4 9】

構成要素 x_1 に対してプラン生成ルール r_1 を実行し、更新可能性のある G ノード (op_1, op_2, \dots など) で繋がっている) を候補 G ノードリスト 3 3 4 へ追加し、ステップ S 3 に戻る（ステップ S 8）。

【0 1 5 0】

以下では、構造化文書データベースへの検索コマンドの具体例を用いて検索グラフの生成から検索プランの生成、実行にわたってより具体的に説明する。

【0 1 5 1】

図 1 9 に、以下で用いる検索コマンド例を示す。この例は、『構造化文書データベース中に出現する「特許」情報に対して、下位の「名称」情報が「検索」という文字列を含んでいるならば、「名称」情報を抽出して「文献」情報として検索せよ』を意味している。

【0 1 5 2】

図 2 0 に、図 1 9 の検索要求に対して検索グラフ生成部 3 3 が生成する検索グラフの一例を示す。

【0 1 5 3】

「\$ 1」変数Gノードは、『指定された文書パス「root」以下で「\$ 2」変数Gノードよりも階層で上位に存在するDノードに対するGノード変数』を示している。

【0 1 5 4】

「\$ 2」変数Gノードは、『「\$ 1」変数Gノードよりも階層で下位に存在し、「特許」要素名称で始まっているDノードに対するGノード変数』を示している。

【0 1 5 5】

「\$ 3」変数Gノードは、『「\$ 2」変数Gノードから見て「特許」要素名称で始まるDノードの子で『名称』要素名称で始まっているDノードに対するGノード変数』を示している。

【0 1 5 6】

「\$ 4」変数Gノードは、『「\$ 3」変数Gノードから見て「名称」要素名称で始まるDノードの子で要素データを指すDノードに対するGノード変数「\$ x 1」を持つDノードに対するGノード変数』を示している。

【0 1 5 7】

「\$ x 1」変数Gノードは、『「\$ 4」変数Gノードから見て要素データを指すDノードで「検索」という文字列を含むDノードに対するGノード変数』を示している。

【0 1 5 8】

このようにGノード同士は2項以上の多項間の制約関係を持っており、それらの変数Dノード群の取りうる値の組み合わせを制約充足的に解くことになる。

【0 1 5 9】

図 2 1 に、本実施形態の検索プラン生成部 3 3 により生成される検索プランの一例を示す。

【0 1 6 0】

図 2 1 の検索プランは、図 1 9 の検索要求を入力とし、図 1 6 のプラン生成ルールを用いた場合の検索プラン生成部 3 3 の出力結果例である。

【0161】

本プラン生成を行うに当たり、構造化文書データベースへの前提として以下のものを想定している。

- ・要素名称生起インデックスファイルの中に「特許」というキーが存在している。
- ・データ生起インデックスファイルの中に「検索」というキーが存在している。

【0162】

検索グラフの全Gノードを候補Gノードリスト334に登録した後、図18のフローチャートにしたがってシミュレーションする。ステップS3からステップS7の1処理を1サイクルとして、変数 `rset` の変化を追ってみると、次のようになる。

【0163】

(第1サイクル)

`rset = {<TAG01, ルール01, 1.0>, <TAG03, ルール03, 0.2>, <CMP01, ルール31, 1.0>, <CMP01, ルール32, 0.1>}`

ここで、`<CMP01, ルール32, 0.1>` が選択され、`FIND` が出力される。

伝播するGノード群は `{ $x1 }` である。

(第2サイクル)

`rset = {<TAG01, ルール01, 1.0>, <TAG03, ルール03, 0.2>}`

ここで、`<TAG03, ルール03, 0.2>` が選択され、`PATHEXPAND2` が出力される。

伝播するGノード群は `{ $__2, $__3 }` である。

【0164】

(第3サイクル)

`rset = {<TAG01, ルール01, 1.0>, <TAG04, ルール0`

2、0.5>、<TAG02、ルール06、0.6>}

ここで、<TAG04、ルール02、0.3>が選択され、PATHEXPAND2が出力される。

伝播するGノード群は {\$_4} である。

【0165】

(第4サイクル)

rset = {<TAG01、ルール01、1.0>、<VAL01、ルール21、0.2>、<TAG02、ルール06、0.6>}

ここで、<VAL01、ルール21、0.2>が選択され、VALUEが出力される。

伝播するGノード群は {\$x1} である。

【0166】

(第5サイクル)

rset = {<TAG01、ルール01、1.0>、<\$x1、ルール11、0.5>、<TAG02、ルール06、0.6>}

ここで、<\$x1、ルール11、0.5>が選択され、JOINが出力される。

伝播するGノード群は {} である。

【0167】

(第6サイクル)

rset = {<TAG01、ルール01、1.0>、<TAG02、ルール06、0.6>}

ここで、<TAG02、ルール06、0.6>が選択され、NOPが出力される。

伝播するGノード群は {} である。

【0168】

(第7サイクル)

rset = {<CON01、ルール71、1.0>}

ここで、<CON01、ルール71、1.0>が選択され、CONSTRUCT

Tが出力される。

【0169】

この実行プランリスト335の意味は、以下のようなものである。

【0170】

(ステップ1)

「検索」という文字列データを含むDノード群を検索する。データ生起インデックスファイルには「検索」というキーが存在しているため、この情報を優先的に利用する。

(ステップ2)

「特許」要素名称を持つ子供Dノード群を取り出す。要素名称生起インデックスファイルの中に「特許」というキーが存在しているため、この情報を優先的に利用する。

(ステップ3)

上記Dノード群で「名称」要素名称を持つ子供Dノード群を取り出す。

(ステップ4)

上記Dノード群で要素データを持つ子供Dノード群を取り出す。

(ステップ5)

ステップ1で検索したDノード群とステップ4で取り出したDノード群の結合(JOIN)を取る。

(ステップ6)

「特許」より上位の文書パスは「root/*」なので、何もしない。

(ステップ7)

上記Dノード群のデータを使って「文献」情報を作り出す。

【0171】

図22に、図21で示した検索プランの実行イメージを示す。

【0172】

ステップ1において、図14のデータ生起インデックスファイルに「検索」というキーが存在するので、直ちに、Dノード群が得られる。

【0173】

一方、ステップ2において、図13の要素名称生起インデックスファイルに「特許」というキーが存在するので、直ちに、\$2と\$3が具体化される。そして、ステップ3において、図5から\$4が具体化される。そして、ステップ4において、要素データを持つDノード群を取り出す。ステップ5において、ステップ1で検索したDノード群とステップ4で取り出したDノード群の結合（JOIN）を取る。

【0174】

このように、インデックスを用いて効率的に検索プランを生成していることがわかる。

【0175】

図23に、図19の検索要求を処理した検索結果の一例を示す。

【0176】

検索結果もXMLで表現され、「文献」情報のリストとして表示されている。「情報検索装置」など「検索」という文字列を含んでいる。

【0177】

ここで、比較のために従来手法でアプローチした場合について説明する。

【0178】

図24に、従来手法でアプローチした場合の検索プランの一例を示す。

【0179】

この従来手法は、ルート要素から始まって、親要素から子供要素群へと展開して、階層的な構造上の構成要素を特定する検索処理方式である。

【0180】

この実行プランリストの意味は、以下のようなものである。

【0181】

（ステップ1）

rootに相当するDノード群を取り出す。

（ステップ2）

上記Dノード群の子孫Dノード群を取り出す。

（ステップ3）

上記Dノード群で「特許」要素名称を持つ子供Dノード群を取り出す。

(ステップ4)

上記Dノード群で「名称」要素名称を持つ子供Dノード群を取り出す。

(ステップ5)

上記Dノード群で要素データを持つ子供Dノード群を取り出す。

(ステップ6)

上記Dノード群でデータが「検索」という文字列データを含むDノード群を選択する。

(ステップ7)

上記Dノード群のデータを使って「文献」情報を作り出す。

【0182】

このように、登録された文書数が増大し階層木の深さと幅が増大すると、検索処理に要する計算量が膨大なものになってしまう。

【0183】

図25に、図24で示した従来手法でアプローチした場合の検索プランの実行イメージを示す。

【0184】

ステップ2での階層木の展開コストが膨大になってしまいうことが容易に想像される。

【0185】

以下では、本実施形態における検索結果をGUI（グラフィカル・ユーザ・インタフェース）的に表示する例について説明する。

【0186】

図26に、図10で示された図9の検索結果をデータ表示用のフィルタープログラムを通してGUI的に表示した一例を示す。

【0187】

「特許」情報に対して、概念「情報モデル」での分類と概念「情報操作」での分類の2分類軸を設定して、「出願番号」と「情報モデル」軸と「情報操作」軸を抽出して「文献」情報として検索した結果であるが、「情報モデル」軸を横軸

に、「情報操作」軸に設定して、2軸のクロスしたデータが「出願番号」情報である。

【0188】

XMLにはスタイルシートという表示フォーマットがあり、XMLドキュメントをWWWブラウザに表示したり、プリンタから印刷したりするときに用いる。スタイルシートの言語として、XSL (Extensible Style Language) が標準の規約として用意されており、これを用いれば図26に示すような情報を出力することができる。

【0189】

図27に、図10で示された図9の検索結果をデータ表示用のフィルタープログラムを通してGUI的に表示した他の例を示す。

【0190】

特許の出願件数を年度別に折れ線グラフ表示させたもので、これも同様に、『特許の出願件数を年度別に集計する』検索要求を処理した結果であるXMLデータに対してスタイルシートを適用すれば、図27に示すような情報を出力することができる。

【0191】

ここで、図1の構造化文書データベース・システムの実現方法のバリエーションについて説明する。

【0192】

本システムは、インターネットもしくはLANなどのネットワークを介して他の計算機から検索要求を受け付け、検索を実行し、ネットワークを介して当該他の計算機に検索結果を返すように実現することも可能である。

【0193】

この場合、他の計算機から図7～図9のような検索要求を受ける代わりに、他の計算機において構文解析し図11／図12のような検索グラフを作成し、これを受けるようにしてもよい。あるいは、図7～図9のような検索要求と図11／図12のような検索グラフのいずれによっても受け付け可能としてもよい。

【0194】

また、要求制御部 1、格納処理部 2、検索処理部 3 を 1 台の計算機上に実装してもよいし、2 台または 3 台の計算機上に別々に実装してもよい。

【0 1 9 5】

また、要求制御部 1、格納処理部 2、検索処理部 3 のそれぞれを実現するプログラムは、記録媒体または通信媒体によって受け渡すことが可能である。この場合、要求制御部 1、格納処理部 2、検索処理部 3 の全てを実現するプログラムを 1 つまたは 1 組の記録媒体に格納して受け渡しすることも、要求制御部 1、格納処理部 2、検索処理部 3 の一部のみを実現するプログラムを 1 つまたは 1 組の記録媒体に格納して受け渡しすることも可能である。

【0 1 9 6】

また、例えば、検索処理部 3 を含むシステムと、格納処理部 2 とデータファイル 4 とインデックスファイル 5 を含むシステムとが、互いに独立したシステムであってもよい。また、検索処理部 3 を含むシステムをサーバとして構成してもよいし、各クライアントに搭載するようにしてもよい。

【0 1 9 7】

もちろん、本システムは、1 つのスタンドアローンのシステムとして実現可能である。

【0 1 9 8】

なお、以上の各機能は、ソフトウェアとしても実現可能である。

【0 1 9 9】

また、本実施形態は、コンピュータに所定の手段を実行させるための（あるいはコンピュータを所定の手段として機能させるための、あるいはコンピュータに所定の機能を実現させるための）プログラムを記録したコンピュータ読取り可能な記録媒体としても実施することもできる。

【0 2 0 0】

本発明は、上述した実施の形態に限定されるものではなく、その技術的範囲において種々変形して実施することができる。

【0 2 0 1】

【発明の効果】

本発明によれば、構造化文書データベースに関する情報を有効に利用しながら、検索要求から生成した検索グラフを最適に巡回することで最適な検索プランを生成し実行することによって、計算量を増大させずに、（曖昧パスを含む）文書が持つ階層構造に対する多様な検索指定による検索を可能にすることができる。

【図面の簡単な説明】

【図 1】

本発明をの一実施形態に係る構造化文書データベース・システムのシステム構成例を示す図

【図 2】

構造化文書の一例を示す図

【図 3】

概念情報の一例を示す図

【図 4】

概念情報の一例を示す図

【図 5】

構造化文書データベースの概念的な構造例を示す図

【図 6】

構造化文書データベースへの構造化文書の格納コマンドの一例を示す図

【図 7】

構造化文書データベースへの検索コマンドの一例を示す図

【図 8】

構造化文書データベースへの検索コマンドの他の例を示す図

【図 9】

構造化文書データベースへの検索コマンドのさらに他の例を示す図

【図 10】

検索要求を処理した検索結果の一例を示す図

【図 11】

検索要求に対して検索グラフ生成部が生成する検索グラフの一例を示す図

【図 12】

検索要求に対して検索グラフ生成部が生成する検索グラフの一例を示す図

【図 1 3】

インデックスファイルの一種である要素名称生起インデックスの概念的な構造例を示す図

【図 1 4】

インデックスファイルの一種であるデータ生起インデックスの概念的な構造例を示す図

【図 1 5】

検索プラン生成部の構成例を示す図

【図 1 6】

検索プラン生成部で利用するプラン生成ルールの一例を示す図

【図 1 7】

検索プラン実行部で利用するオペレータの一例を示す図

【図 1 8】

検索プラン生成部の処理手順の一例を示すフローチャート

【図 1 9】

構造化文書データベースへの検索コマンドのさらに他の例を示す図

【図 2 0】

検索要求に対して検索グラフ生成部が生成する検索グラフの他の例を示す図

【図 2 1】

検索プラン生成部により生成された検索プランの一例を示す図

【図 2 2】

検索プランの実行イメージを示す図

【図 2 3】

検索要求を処理した検索結果の他の例を示す図

【図 2 4】

従来手法でアプローチした場合の検索プランを示す図

【図 2 5】

従来手法でアプローチした場合の検索プランの実行イメージを示す図

【図 2 6】

検索結果をデータ表示用のフィルタープログラムを通して G U I 的に表示した一例を示す図

【図 2 7】

検索結果をデータ表示用のフィルタープログラムを通して G U I 的に表示した他の例を示す図

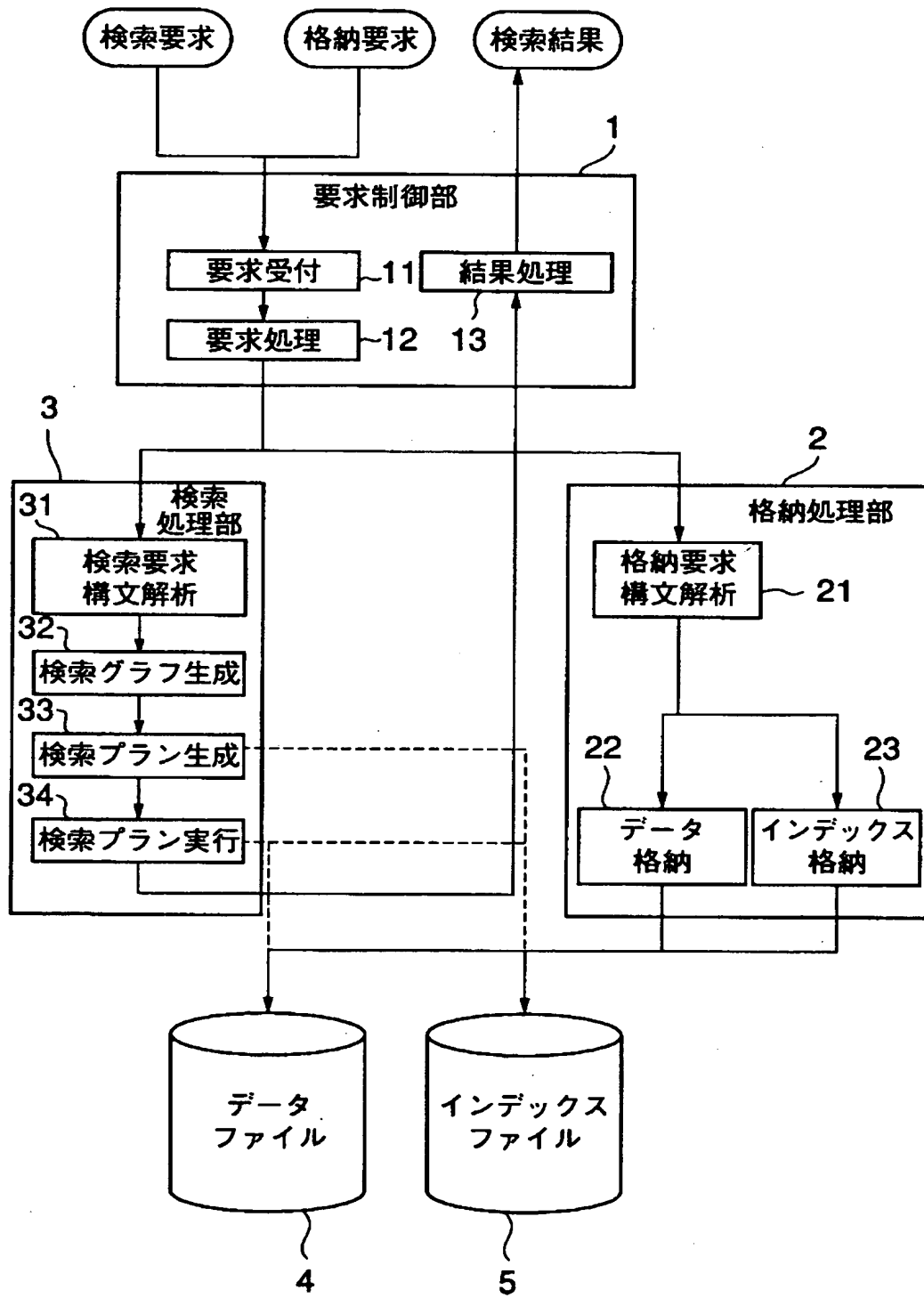
【符号の説明】

- 1 … 要求制御部
- 2 … 格納処理部
- 3 … 検索処理部
- 4 … データファイル
- 5 … インデックスファイル
- 6 … プラン生成ルール
- 1 1 … 要求受付部
- 1 2 … 要求処理部
- 1 3 … 結果処理部
- 2 1 … 格納要求構文解析部
- 2 2 … データ格納部
- 2 3 … インデックス格納部
- 3 1 … 検索要求構文解析部
- 3 2 … 検索グラフ生成部
- 3 3 … 検索プラン生成部
- 3 4 … 検索プラン実行部

【書類名】

図面

【図 1】



【図 2】

<特許>
<名称>情報検索装置</名称>
<出願人>T社</出願人>
<出願番号>特願平10-XXXXXX</出願番号>
<出願日>
<年>10</年><月>3</月><日>12</日>
<出願日>
<要約>
情報の提示形式の変更が利用者側の観点で自由に行え、情報活用の範囲が広がるとともに、情報活用の促進が図れるデータベースを提供する。
</要約>
<キーワード>XML</キーワード>
<キーワード>検索</キーワード>
</特許>

【図 3】

```

<概念名前= “情報モデル” >
  <概念名前= “ドキュメント” >
    <概念名前= “構造化ドキュメント” >
      <概念名前= “XML” />
      <概念名前= “SGML” />
    </概念>
    <概念名前= “非構造化ドキュメント” >
      <概念名前= “テキスト” />
    </概念>
  </概念>
  <概念名前= “リレーション” >
    .....
  </概念>
  <概念名前= “オブジェクト” >
    .....
  </概念>
</概念>

```

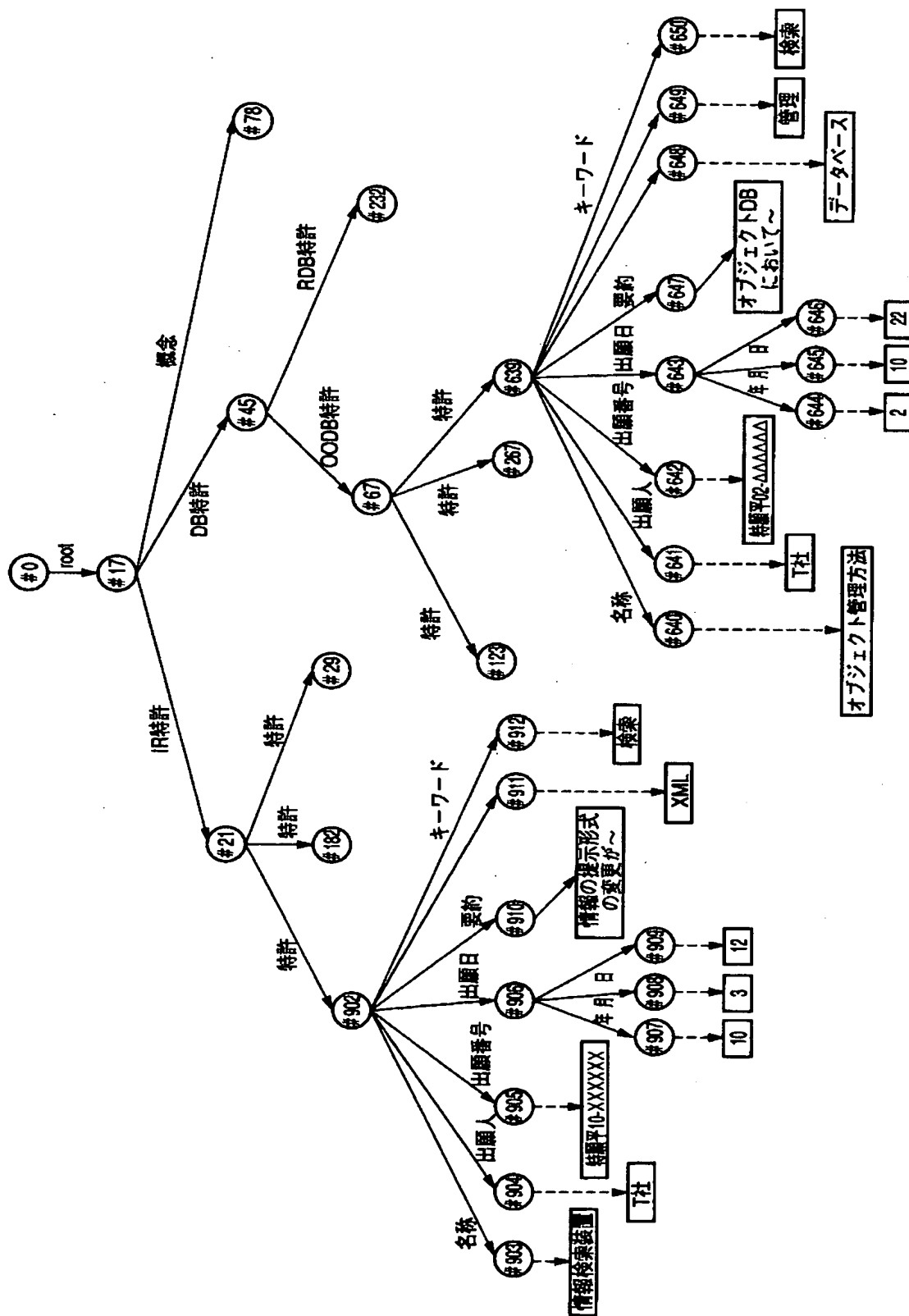
【図 4】

```

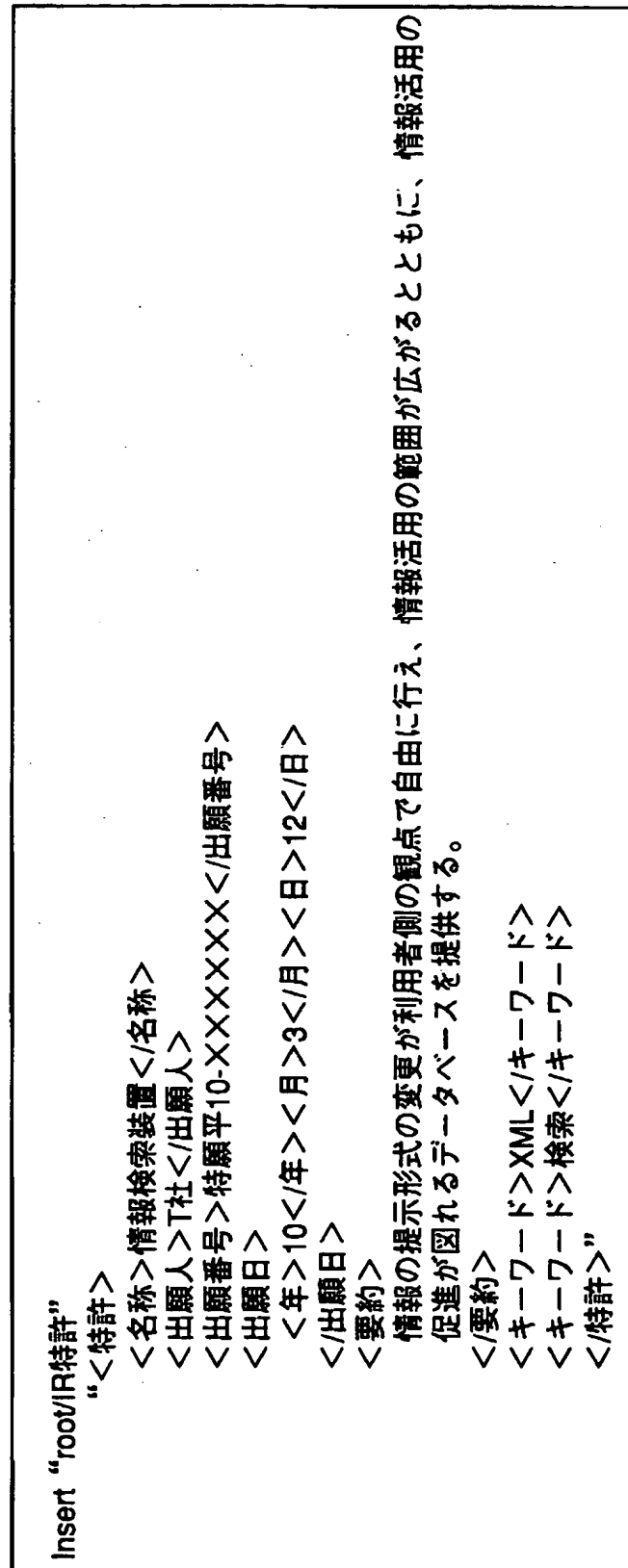
<概念名前= “情報操作” >
  <概念名前= “検索” />
  <概念名前= “格納” />
  <概念名前= “加工” />
  <概念名前= “流通” />
</概念>

```

【図 5】



【図 6】



【図 7】

```

Select
  <文献>
    <出願番号>$x1</出願番号>
  </文献>
Where
  <*/特許>
    <出願番号>$x1</出願番号>
    <キーワード>$x2</キーワード>
  </特許>From "root/"
  $x2= "検索"

```

【図 8】

```

Select
  <文献>
    <出願番号>$x1</出願番号>
  </文献>
Where
  <*/特許>
    <出願番号>$x1</出願番号>
    <キーワード>$x2</キーワード>
  </特許>From "root/"
  <概念名前= "ドキュメント" >
    <*/概念名前=$x2/>
  </概念>From "root/"

```

【図 9】

```

Select
  <文献>
    <出願番号>$x1</出願番号>
    <分類軸=“情報モデル”>$3</分類>
    <分類軸=“情報操作”>$4</分類>
  </文献>
Where
  <*/特許>
    <出願番号>$x1</出願番号>
    <キーワード>$x2</キーワード>
  </特許>From “root/”
  <*/概念名前=“情報モデル”>
    <概念名前=$3>
    <*/概念名前=$x2/>
  </概念>
  </概念>From “root/”
  <*/概念名前=“情報操作”>
    <概念名前=$4>
    <*/概念名前=$x2/>
  </概念>
  </概念>From “root/”

```

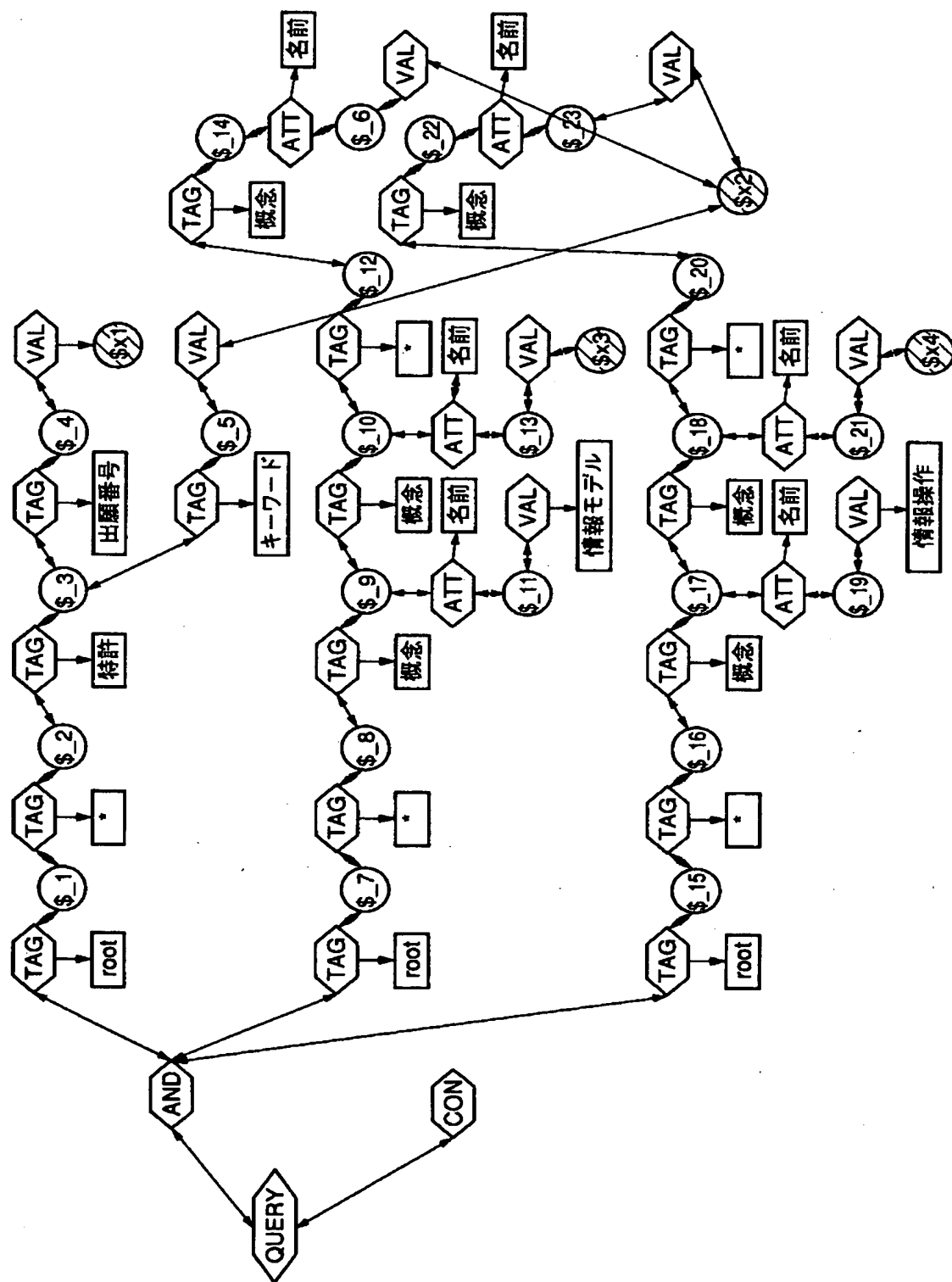
【図 1 0】

```

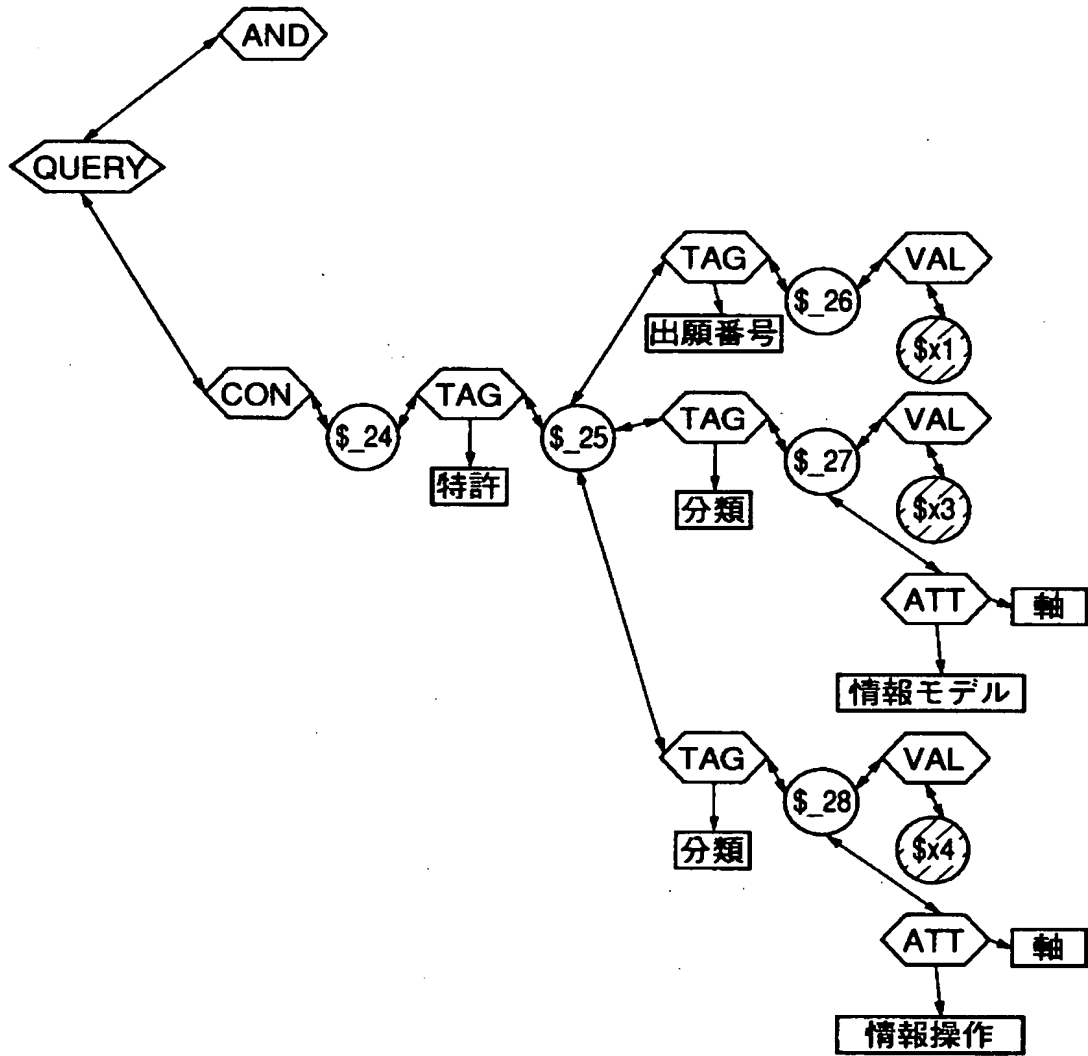
<結果>
  <文献>
    <出願番号>特願平10-XXXXXX</出願番号>
    <分類軸=“情報モデル”>ドキュメント</分類>
    <分類軸=“情報操作”>検索</分類>
  </文献>
  <文献>
    <出願番号>特願平09-□□□□□□</出願番号>
    <分類軸=“情報モデル”>リレーション</分類>
    <分類軸=“情報操作”>格納</分類>
  </文献>
  <文献>
    <出願番号>特願平10-000000</出願番号>
    <分類軸=“情報モデル”>リレーション</分類>
    <分類軸=“情報操作”>検索</分類>
  </文献>
</結果>

```

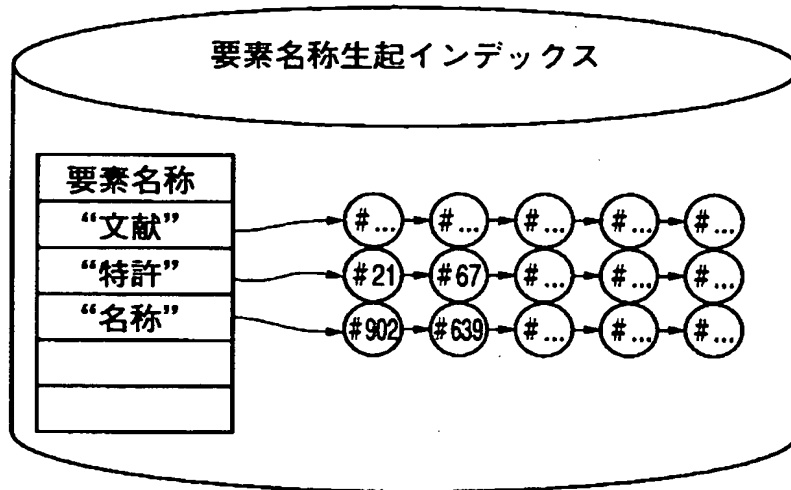
【図 1 1】



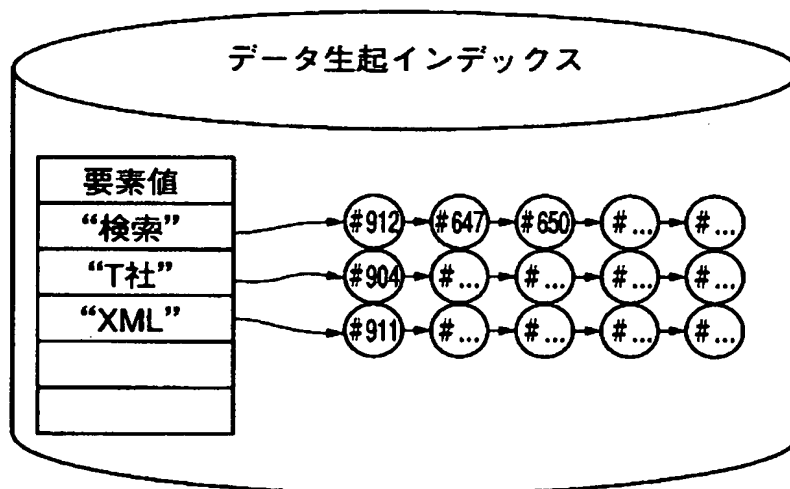
【図 1 2】



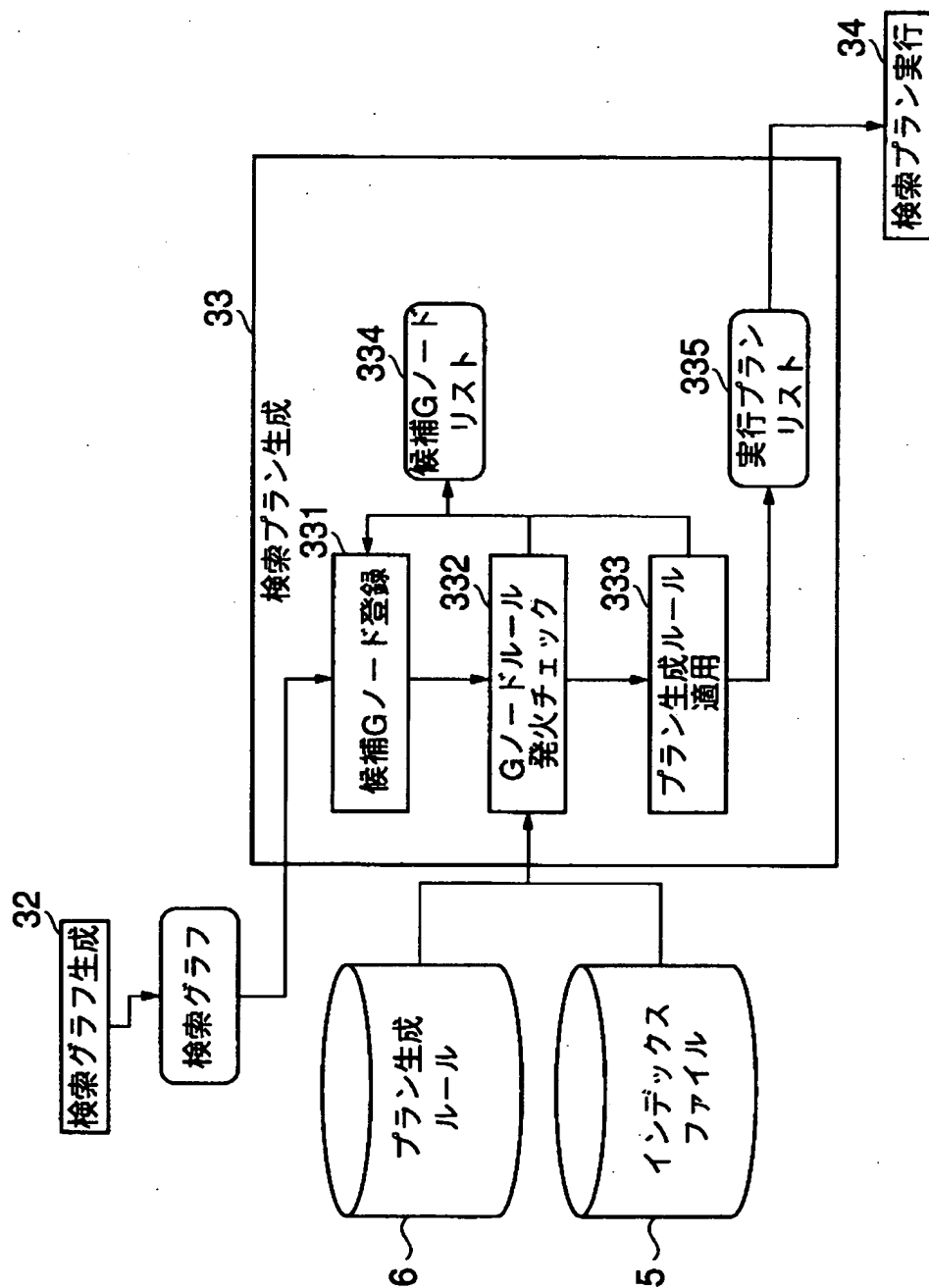
【図 1 3】



【図 1 4】



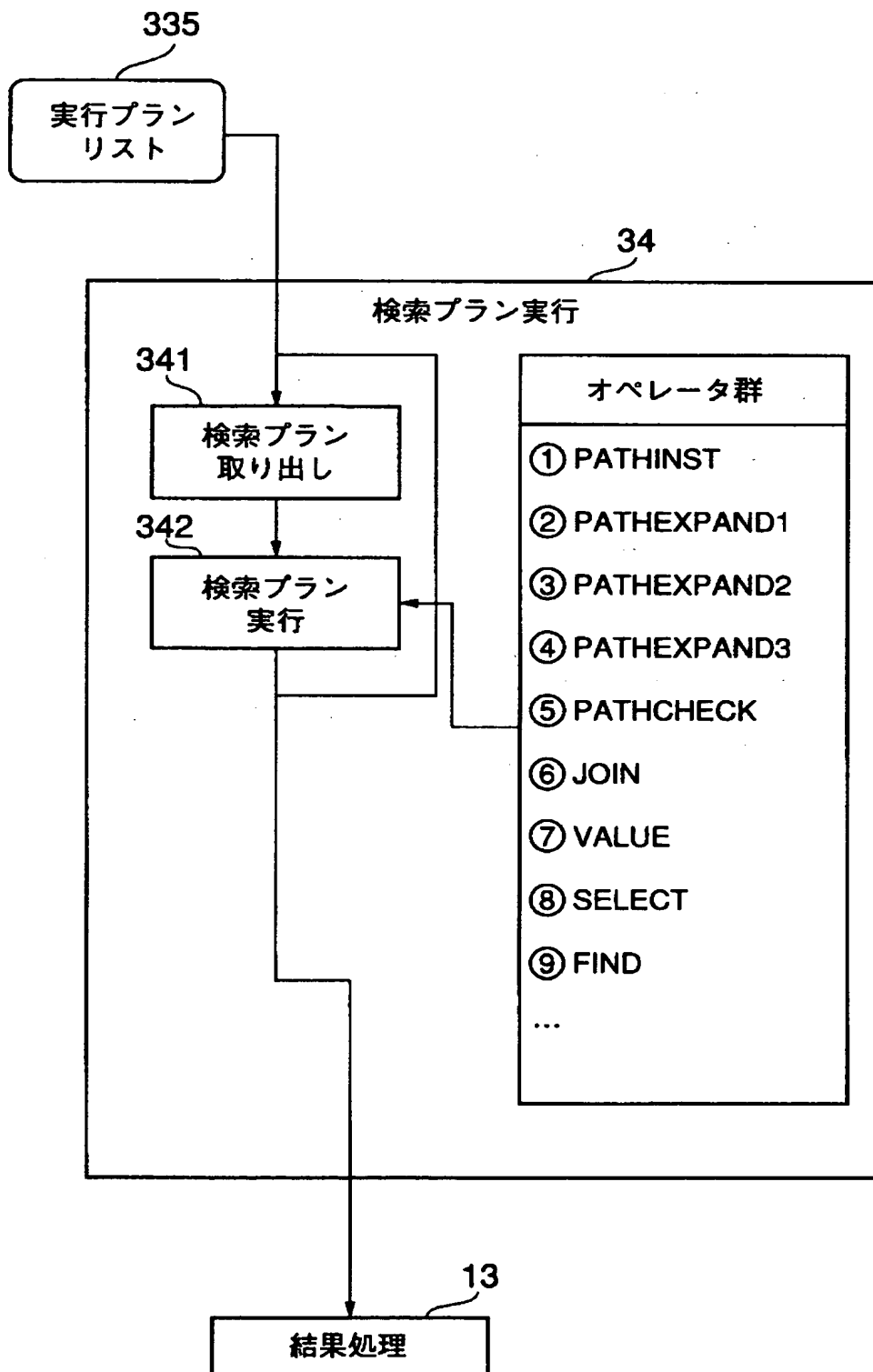
【図 1 5】



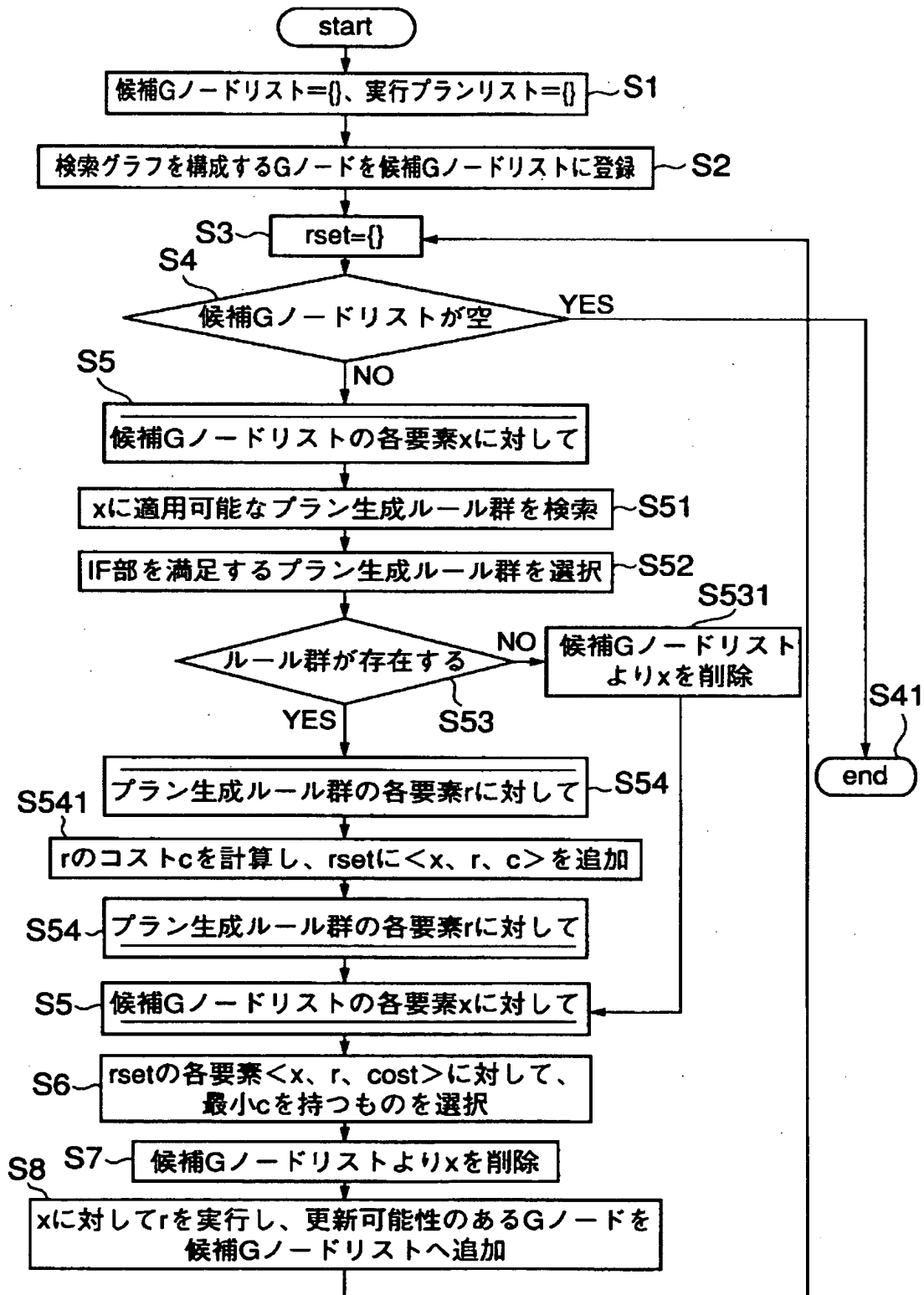
【図 1 6】

番号	Gノード	コスト	IF				THEN
			OP1	OP2	OP3	その他	
01	TAG TAG TAG	1.0	AND	具	未	OP2に要素名称生起 インデックスが存在	PATHINST PATHEXPAND1 PATHEXPAND2
02		0.5	具	具	未		
03		0.2	未	具	具		
04	TAG	0.1	未	具	具	OP3に対する逆階層 インデックスが存在	PATHEXPAND3
05	TAG TAG	0.3	具	具	具	OP1の隣接TAGノード のOP2が "root"	PATHCHECK NOP
06		0.6	具	具	具		
...							
11	VAR	0.5	具	具	具		JOIN
...							
21	VAL	0.2	具	未	未		VALUE
...							
31	CMP CMP	1.0	具	未	未	OP2に対するデータ インデックスが存在	SELECT FIND
32		0.1	具	具	具		
...							
41	AND						
...							

【図 17】



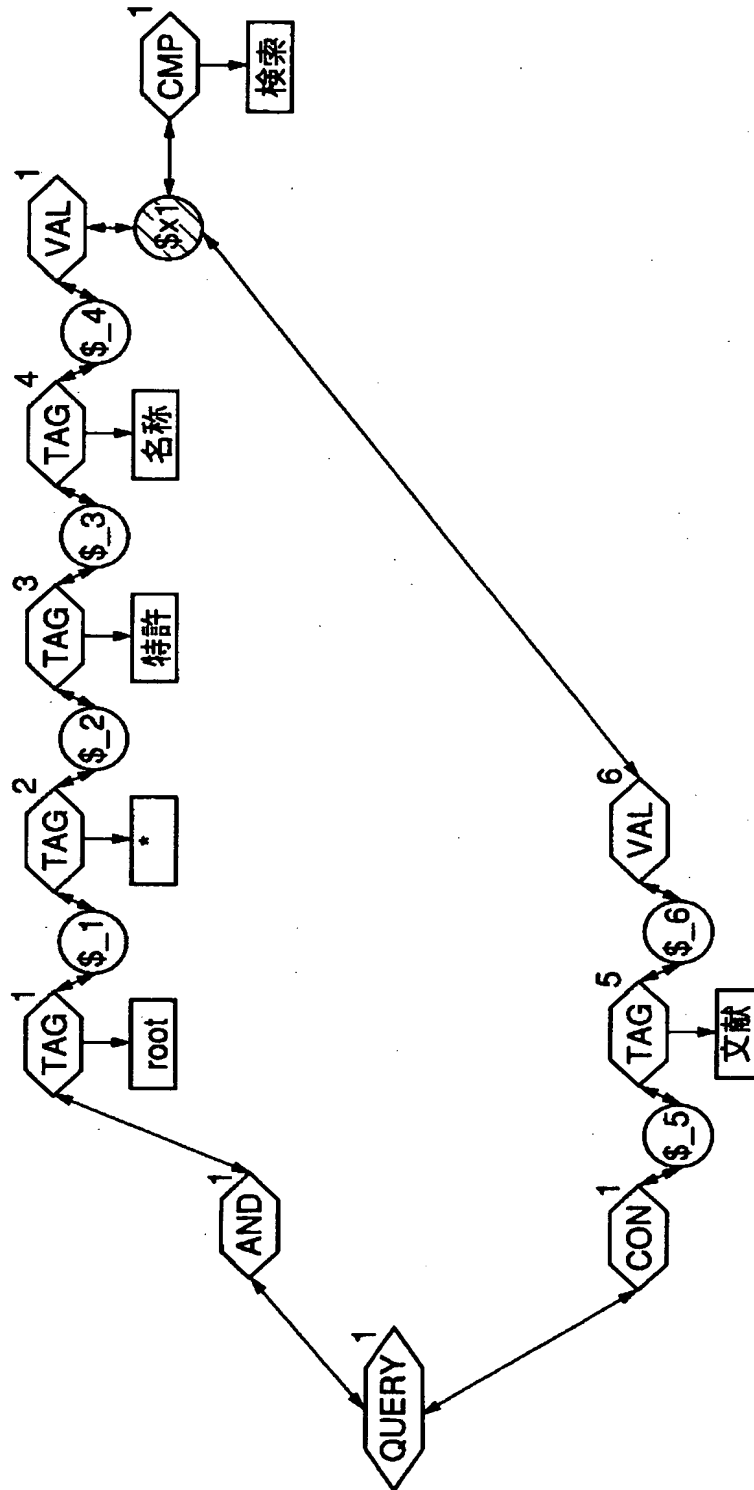
【図 1 8】



【図 1 9】

```
Select
<文献>$x</文献>
Where
  <*/特許>
    <名称>$x</名称>
  </特許> From "root/"
  $x like "検索"
```

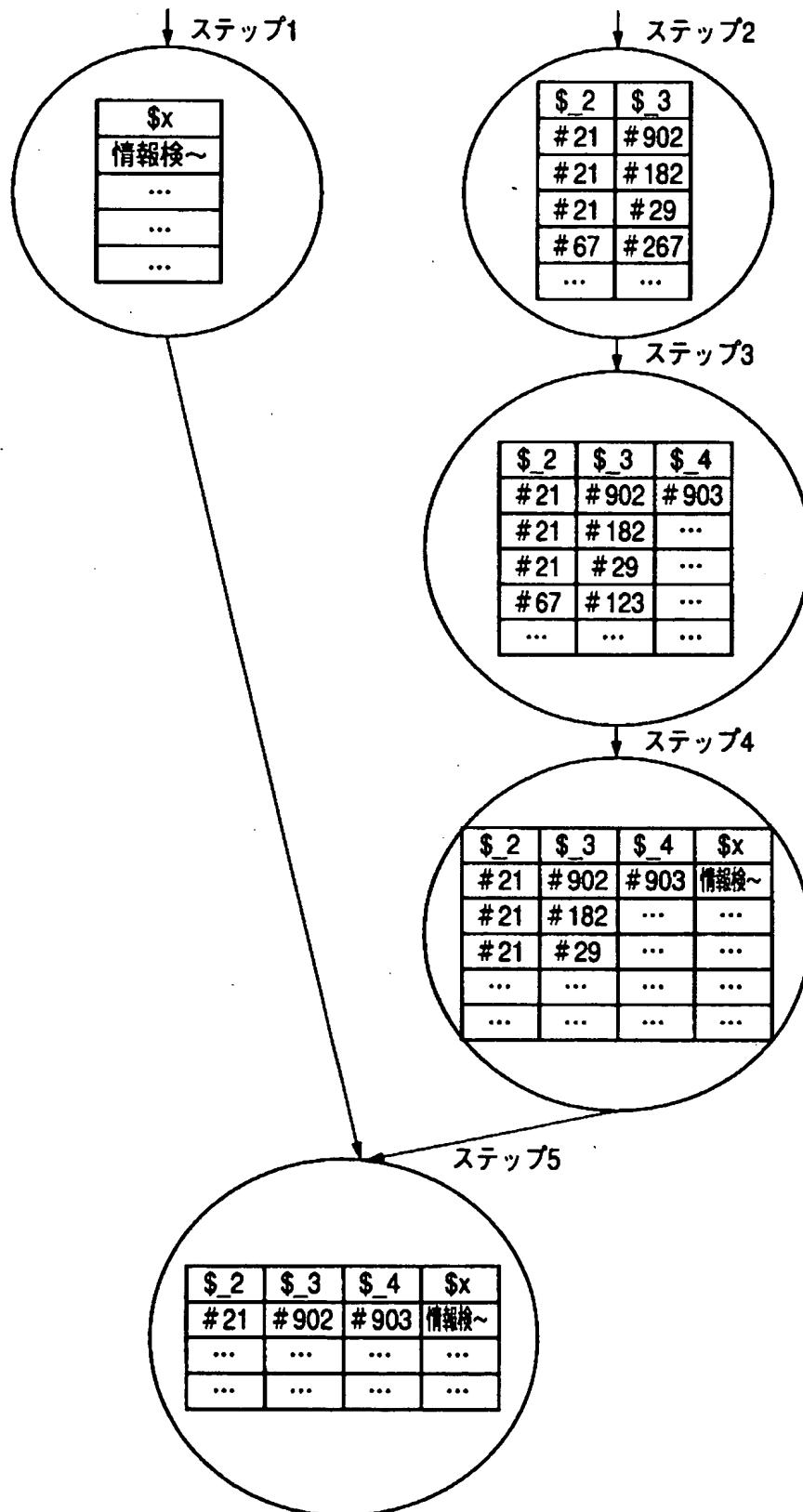
【図 2 0】



【図 2 1】

ステップ	G ノード	ルール番号	オペレータ
ステップ1	CMP	ルール32適用	FIND
ステップ2	TAG	ルール03適用	PATHEXPAND2
ステップ3	TAG	ルール02適用	PATHEXPAND1
ステップ4	VAL	ルール21適用	VALUE
ステップ5	VAR	ルール11適用	JOIN
ステップ6	TAG	ルール06適用	NOP
ステップ7	CON	ルール適用	CONSTRUCT

【図 2 2】



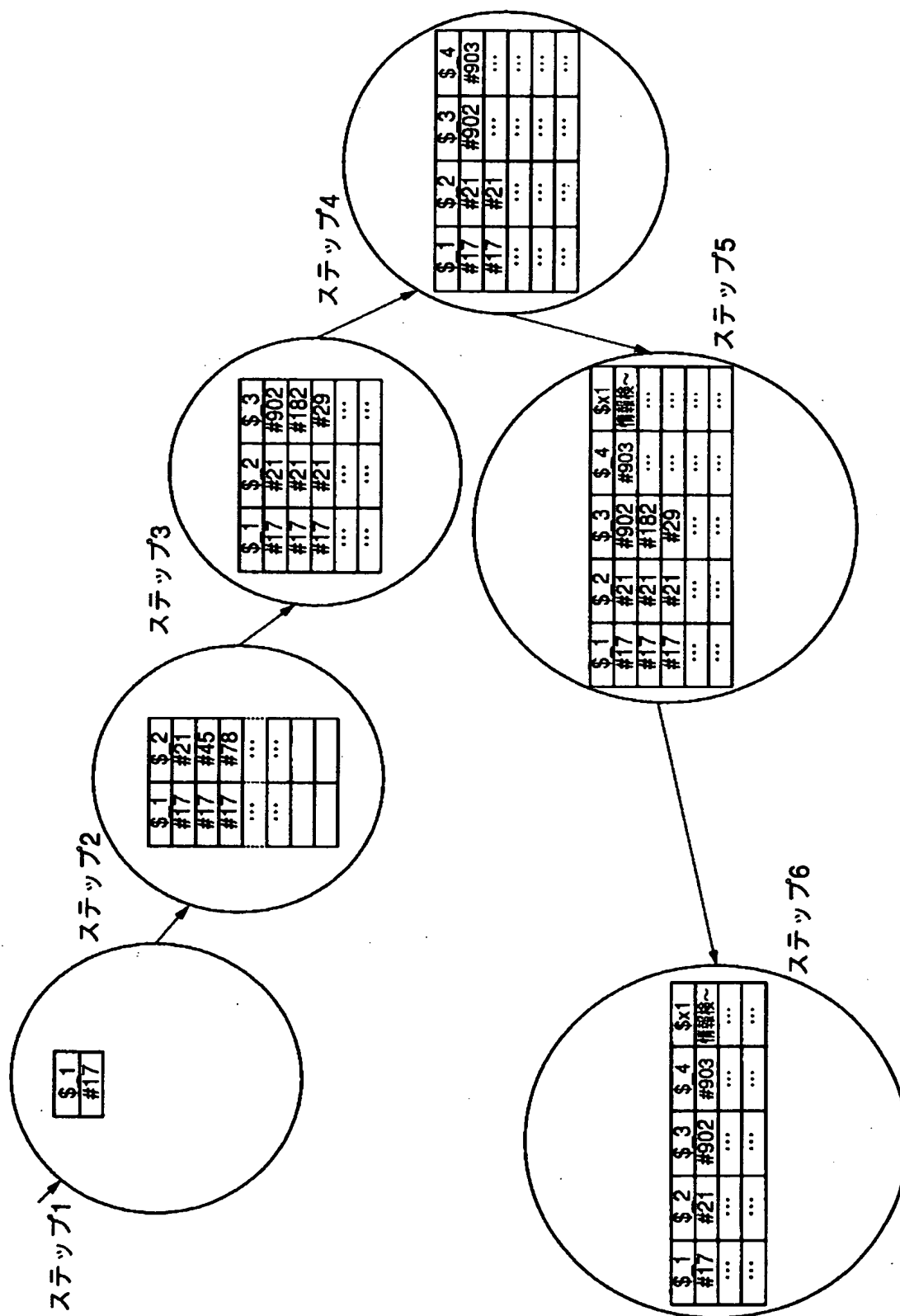
【図 2 3】

<結果>
 <文献>情報検索装置</文献>
 <文献>データ検索方式</文献>
 <文献>検索処理の最適化の方法と検索装置</文献>
 </結果>

【図 2 4】

ステップ	G ノード	ルール番号	オペレータ
ステップ1	TAG	ルール01適用	PATHINST
ステップ2	TAG	ルール02適用	PATHEXPAND1
ステップ3	TAG	ルール02適用	PATHEXPAND1
ステップ4	TAG	ルール02適用	PATHEXPAND1
ステップ5	VAL	ルール21適用	VALUE
ステップ6	CMP	ルール31適用	SELECT
ステップ7	CON	ルール適用	CONSTRUCT

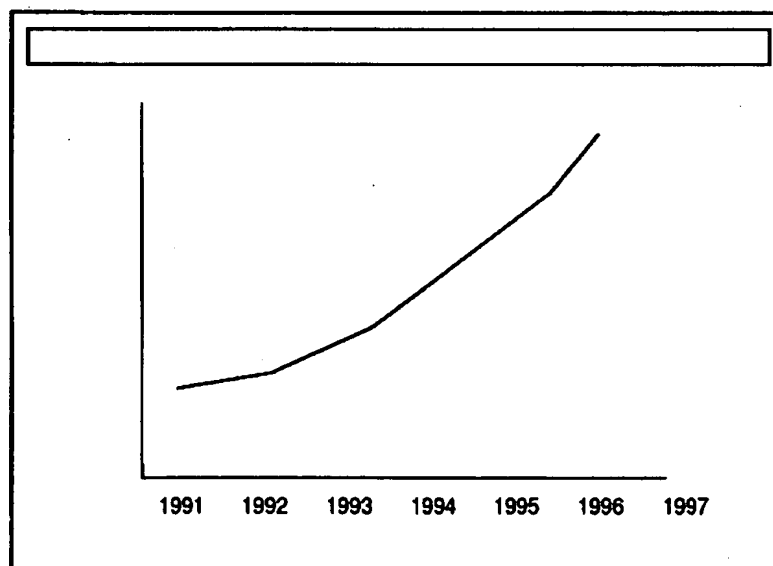
【図 25】



【図 2 6】

	ドキュメント	リレーション	オブジェクト
検索	特願平10-XXXXXX	特願平10-000000	
格納		特願平09-000000	
加工			
流通			

【図 2 7】



【書類名】 要約書

【要約】

【課題】 計算量を増大させずに、曖昧パスを含む文書が持つ階層構造に対する多様な検索指定を行うことを可能とした、構造化文書検索方法を提供すること。

【解決手段】 文書の論理構造を含む検索要求に基づいて文書の構造情報を含む検索グラフを生成し（3 1, 3 2）、データの生起位置を特定するデータ生起インデックスと要素名称の生起位置を特定する要素名称生起インデックスを含むインデックスファイル 5 を利用しながら検索グラフ中において評価可能な部分グラフを優先的に評価する戦略に基づいて検索グラフを巡回することによって最適な検索プランを生成し（3 3）、データファイル 4 に対して検索プランを実行することによって（3 4）、検索要求を満足する検索結果を求める。

【選択図】 図 1

出 願 人 履 歴 情 報

識別番号 [000003078]

1. 変更年月日	1990年 8月22日
[変更理由]	新規登録
住 所	神奈川県川崎市幸区堀川町72番地
氏 名	株式会社東芝